SELF-ORGANIZING CONFERENCE ON MACHINE LEARNING 2018

A PREPRINT

SOCML 2018 Note Compilers Team on Behalf of SOCML Moderators and Note-takers *†‡§

March 12, 2019

ABSTRACT

This is a summary of notes taken during the Self Organizing Conference on Machine Learning (SOCML), held at the Google Toronto office on November 30 and December 1, 2018. The conference was co-organized by Ian Goodfellow, Geoffrey Hinton and the University Relations team at Google

***K*eywords** Social Conference, Machine Learning, Deep Learning, Reinforcement Learning

Self Organizing Conference in Machine Learning (SOCML) is an "unconference" where attendees drive the session agenda. Multiple hour-long sessions run simultaneously on topics that are nominated by the SOCML participants . Attendees come from diverse backgrounds with different levels of expertise in machine learning and have focused discussions about the session topics. Participants also volunteer to become session moderators and note-takers. ***List of Abbreviations***

AI	Artificial Intelligence
DL	Deep Learning
GAN	Generative Adversarial Networks
ML	Machine Learning
RL	Reinforcement Learning
SOCML	Self-Organizing Conference on Machine Learning
TF	Tensorflow
VAE	Variational Autoencoders
DandI	Diversity and Inclusion
GDPR	General Data Protection and Regulation
HTM	Hierarchical Temporal Memory

*Meltem Atay, Department of Neuroscience and Neurotechnology, Middle East Technical University,Turkey, Ankara, meltem.atay@metu.edu.tr

†Khimya Khetarpal, McGill University,School of Computer Science,Montreal, Quebec, khimya.khetarpal@mail.mcgill.ca

‡Miti Modi, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario; Applied Machine Learning Group, Loblaw Companies Limited, Brampton, Ontario, miti.modi@mail.utoronto.ca

§Cynthia Habonimana, Department of Computer Engineering and Computer Science, California State University, Long Beach, California; cyn.habonimana@gmail.com

		Room name	Kawarthas (65)	Algonquin (170), Split 1*	Algonquin (170), Split 2*	Georgian Bay (20)*	Muskoka (13)		
		Capacity	65	50	50	20	13		
				*video not supported	*video not supported	*video not supported			
Day 1	Time 10-11:00 AM	**Session 1 Topic**	**RL in Real Life**	**Adversarial examples, fraud and abuse detection**	**AI Education**	**Autonomous Vehicles**	**Time series modeling / forecasting**		
		Session 1 Moderator	Yuxi Li	Nicolas Papernot	Meltem Atay	Hager Radi	Bongba Ange Patrick Kakou		
		Session 1 note-taker	Jessy Lin	Nikolaos Sarafianos	Prashnna K Gyawali	Khimya Khetarpal	Odd Sandbekkhaug		
	Time 1-2:00 PM	**Session 2 Topic**	**Reducing the dependence on massive labeled datasets**	**Language modeling + machine translation**	**Justifiability, Explainability, Interpretability and Algorithmic Decisions**	**ML for Creativity**	**AI governance / national AI strategies**		
		Session 2 Moderator	Colin Raffel	Judy Hanwen Shen	Gillian Hadfield	Tom White	Bogdana Rakova		
		Session 2 note-taker	Jörn Jacobsen	Dustin Tran	David Madras	David Kale			
	Time 2:45-3:45 PM	**Session 3 Topic**	**AI for Social Good**	**Interesting, hard to explain phenomena**	**ML in Supply Chain**	**Curiosity-Driven RL**	**Technical AI Safety**		
		Session 3 Moderator	Bogdana Rakova		Miti Modi	Khimya Khetarpal	Victoria Krakovna + David Krueger		
		Session 3 note-taker	Nikolaos Sarafianos			Erin Grant	Tegan Maharaj		
		Room name	Kawarthas (65)	Algonquin (170), Split 1*	Algonquin (170), Split 2*	Temagami (28)	Georgian Bay (20)*	Muskoka (13)	
		Capacity	65	50	50	28	20	13	
				*video not supported	*video not supported		*video not supported		
Day 2	Time 10-11:00 AM	**Session 1 Topic**	**ML for Healthcare / Medical Imaging in TF**	**Neuroscience-inspired ML**	**ML for Climate**	**Diversity and Inclusion**	**Inverse RL**		
		Session 1 Moderator	Marzyeh Ghassemi	Simon Kornblith	Soukayna Mouatadid	Cody Coleman			
		Session 1 note-taker	Neil Tenenholtz	Meltem Atay	Tegan Maharaj				
	Time 1-2:00 PM	**Session 2 Topic**	**Generative Models for RL**	**Causal Inference, especially for small-data regime**	**Strategies to Reduce the Computational Cost**	**Ethical / responsible AI development**	**Building and monitoring production-ready ML / distributed training**		
		Session 2 Moderator	Dustin Tran		Murium Iqbal	Tim Hwang	Cody Coleman		
		Session 2 note-taker	Erin Grant	Jessy Lin	Hager Radi	Parinaz Sobhani			
	Time 2:45-3:45 PM	**Session 3 Topic**	**AGI / Alternatives to the Reward Maximization Framework**	**Building successful AI startups**	**Reproducibility**	**Role of ML in non-tech industries**	**Differential Privacy**	**Limitations of the GAN framework**	
		Session 3 Moderator	Steven Stenberg Hansen	Parinaz Sobhani	Gideon Dresdner		Nicolas Papernot	Alexia Jolicoeur-Martineau	
		Session 3 note-taker	Ashley Edwards		Khimya Khetarpal				

Figure 1: SOCML Tentative Schedule socml.org

1 DAY 1

The SOCML program is shown in Figure **??**. 16 discussion sessions were held on Day 1.

1.1 Day 1 Session 1

Session 1 was scheduled between 10-11 am. The session topics were *RL in Real Life*, *Adversarial Examples, Fraud and Abuse Detection*, *AI Education*, *Autonomous Vehicles*, *Time Series Modelling/Forecasting*, and *Disentangling*.

1.1.1 RL in Real Life

Session Moderator: **Yuxi Li**, University of Alberta
Session Note-taker: **Jessy Lin**, Google Research & MIT Computational Cognitive Science Lab

Brief information about the session: There were 35 attendees in that session. Several topics were discussed including a general overview of RL, challenges of applying RL in practice and issues with deployment of a naive RL agent.

Main discussion points:

Overview of RL in real life and key challenges: Massive computing and deep reinforcement learning have helped build systems that can play games such as Go successfully. Modern applications of RL include Facebook Horizon ML, Google AutoML data center cooling and Amazon SageMaker. LiLi [2018]. In this session the main topics explored was the application of RL in real life discussing the following questions:

- How can real life applications handle the sparse and delayed feedback from the formulation in RL?
- How can sample-hungry RL algorithms be used in real life? Are there any sample efficient algorithms in RL?
- What are the consequences of building an imperfect model?
- What is the current landscape for transferring simulation results to the real world?
- How should the performance of applied-RL model be validated?
- How is the impact of modelling errors estimated?
- Consideration for ethics in RL, especially if used to build "killer" applications?

Most of the success in RL has evidently been in simulators such as Atari, Open AI Gym, etc. where over-fitting on a specific game or a task still continues. Some of the biggest barriers in deploying a successful RL system include:

- Availability of large compute.
- Noisy data makes it challenging to build RL systems that provide continuity.
- Model performance evaluation. There are examples where RL algorithms have replaced rule-based systems for conversational/dialogue agents cite. However, there are concerns about value alignment and if the agents actually produce desired outputs in the real world.
- There can be promising scientific applications of RL in the fields of bio-medical sensing, medical intervention, synthesis pathways for molecules and materials. However, definitions and limitations of real world are vague and applications outside the paradigm of agent navigating in environment (not games and robotics, but health-care, energy, etc.) needs to be defined more formally and precisely.

How to deploy a naive RL agent in the real world if it can decide to take random actions?

Is AutoML a solution? AutoML offers an interesting intersection of classical control (MPC) and RL. AutoML is initiated by learning from logs and proposing suggestions with a human-in-the-loop. WhatâĂŹs the advantage of AutoML anyway over classical control algorithms? — Looking for non-obvious solutions than domain experts. First cases of RL in real life will probably be areas where there are already active control algorithms, so we have a baseline. However, we still need a human specifying the reward signal. This process is often time-consuming and error-prone. (What would you do otherwise? See session tomorrow!). cite the session these notes talk about. Another aspect of this is Safety problems e.g. killing everyone to reduce sickness counters factual policy evaluation? In addition, all the RL algorithms require exploratory actions, which is ridiculous in critical applications such as health-care. If youâĂŹre exploring, youâĂŹre not doing what is best right now. Multi-objective rewards and policies seem very promising off-policy, but when the behavior of the model does not confirm prior expectations of domain experts, people still choose not to follow that policy. We are currently not in a place to replace humans He et al. [2018].

Much of the work requires *interpretations and explainability*

Positive explanation AND negative explanation, why the humanâĂŹs hypothesis is wrong). These could be posed as social/organizational problems, but they can possibly be framed as research problems at the same time. Really hard to pull that back in to a simulator. If you had a highly detailed physical model and could have that explainability, would that have been sufficient to change behavior? Domain expert intuition that we canâĂŹt explain (maybe visceral understanding of the system, like âĂIJthis piece of hardware craps out, needs be to operated in this specific wayâĂİ). There are always such concerns about factors that are not captured by the model that the humans viscerally understand (e.g. hardware, people in the building, etc.) Maybe the problem is about thinking of it as a replacement instead of a tool or aid?

How about incorporating explainability into the reward signal? The problem is that the decisions are at a timescale you donâĂŹt have enough data (e.g. once-a-day basis). Another example is optimization for antennas. These are seen as non viable designs by some people, while others feel there is something to learn from it. Our goal should be to produce classes of solutions that inform experts, not just to solve the problem. This is a key distinction from games and Atari. In real world problems, domain expert is necessarily going to throw away a lot of the requirements and factors under consideration to just explain the problem simply so lots of different policies are necessary. These problems could be solved by focusing on developing algorithms (that work well in simulation / games, etc.) and focusing on problem formulation (the environment!, not the agent).

How do probabilistic confidence intervals translate to safety factors in traditional engineering? Consider the self driving car application. Are disengagements per mile in autonomous vehicles a good measure of quality? Is deploying RL in real life would be introducing a new kind of problem for regulatory mechanisms ?

Off-policy RL A lot of knowledge we have about the human body is not through interventions which would have been wildly unethical, but from freak accidents, things that have happened in the world, etc. Equivalents in off-policy RL - trajectories that have happened but not based on defined policy. How can we leverage this better especially for real life applications of RL?

1.1.2 Adversarial Examples Fraud and Abuse Detection

Session Moderator: **Nicolas Papernot**, Google Brain
Session Note-taker: **Nikolaos Sarafianos**, University of Houston

There were more than 40 attendees for this session. The main topics discussed were — Real word adversary (messing with your system), WhoâĂŹs winning the arms race?, Usability in content moderation? (detecting fraud/harassment/toxicity).

Story about real world: Detect out of policy things. 1) Catch things by looking at multi modal inputs (image only or text only may not be enough). 2) Images where you want to detect counterfeit. They are easy to do and hard to achieve. Adversarial examples can be used to avoid censorship.

Different teams will be competing between each other. Use adversarial practices to prove something does not work. Different teams that are working on the same problem by finding adversarial examples in other people's project.

Key Ideas Discussed

- For instance, in the scenario of an *advertising company*: Pay the company to display ads. You want to build an add that needs to be clicked. For this new campaign I need data (stop and start new campaign) every time. Prevent the model from learning that itâĂŹs bad by creating a new one. You still want to display as many adds as possible even if theyâĂŹre bad. Adversarial examples can be useful for *Ad Blocking*. They may not be detected by image segmentation. Usually motivations are not clear in such scenarios.
- In the scenario of self driving, goodness of the system versus the responsibility could be taken by insurance company can be evaluated in the adversarial domain. So human passenger should not be normally penalized when there is a system failure during autonomous driving. In this system while processing data, it may also be an option to switch to airplane mode. Another scenario may involve situations under the terms of *White hat hacking*. Intercepted signals between airplanes to tower may not be encrypted or it does not require any encryption. It would be possible to send packets while mimicking the data from airplane.
- Coordinated campaigns to recruit people for radical movements (platforms to avoid words that can be detected). Find less obvious ways to look like an ideology or movement instead. Reverse engineer the system to communicate. There are psychological aspects of adversarial behavior: Think of manipulating human opinions -> bubble effect. It is not clear how to transfer adversarial behavior from ML to humans. it can be difficult to reason about it.
- Public relation companies that create content as an adversary to political opinions? Hard to find. Companies paid to have a topic âĂIJcorruption from white vs black peopleâĂİ. Dealing with fake content being spread although the adversary has disappeared as a company any more.
- Access to training set (rubbish shifts the model). Model learns the wrong thing (poisoning attacks). One direction is to find words that can bypass spam detection. Influence function to understand each samples contribution then mess with it so as dogs can be classified as fish Koh and Liang [2017].
- Human-centered content moderated systems vs community moderator as part of the job they investigate the context and inform the system. For an improvised element hard to automatically feed to the system. A few open questions here are as follows: How do you deliver information to understand how things are moderated and how ML comes in? How do you identify relevant information so as to decide whether to do a behavior after or not? How do we get a good understanding of the behavior of an ML model. How do we identify its properties (robustness/fairness)? How much of logic of the model was exercised?
- In question answering systems: Context is important-> how do you train the model to look at different parts of the system. How much of the adversary examples broke the model? One could may be use ethnographic work to find a series of indicators to find misinformation. Research to identify the indicators. The idea behind *timeline ranking* is that it is easier to interpret a simple model but not when you add additional modules to the model. The more modalities the harder it is to deal with adversarial behavior.

Thoughts on the intersection of interpretability and adversarial examples

- Adversarial robust model is easier to interpret. At MNIST it learns better weights that look like strokes. It is only adversarial because we cannot explain it.
- Even if we get to the point that the models are doing everything correctly there still might be different ways at looking at images (different features from humans vs features from ML).

- Interpretability from normal vs adversarial example might be different.
- Saliency maps can be possible to manipulate. It is not just the noise that causes the mis-classification there are other co-linearities.
- It is difficult to understand if interpretability is possible. Is the logic behind ML necessary close to the logic behind a humanâĂŹs way of thinking?
- Look at distribution of samples that a human failed and the distribution of samples that an ML failed. Different people might fail differently?
- Goal of interpretability? Human can predict how the system will behave. If the human is allowed to choose a particular input we should anticipate how the model will behave on those inputs. When human says input is ambiguous then this could be aligned with the machine.
- Interpretability is domain specific. Does not have to be fully predictable. How different inputs might affect different outputs. What might a user want to do to change the output.
- Interpretability is subjective. How do you detect if an account is fraudulent? How to make the client understand what is interpretable
- Is interpretability model independent or not?
- Interpretability is not one thing. Users must be able to trust decisions, peak the behavior of the system. Developers of the system also should understand its properties.
- ThereâĂŹs a legal definition to explain how you arrived into a decision.

1.1.3 AI Education

Session Moderator: **Meltem Atay**, Middle East Technical University
Session Note-taker: **Prashnna K Gyawali**, Rochester Institute of Technology

There were 15 attendees for this session. Mainly discussed about how AI has recently seen a global traction and thus the AI education has also seen increased attraction, and key topics were human learning, AI education, Policy of AI education in school, AI educator.

AI education

There are many online materials but it can be very confusing for a new learner to try to find the route and get advanced through the steps. Difference between AI education in university vs. online materials can be so much to handle for a beginner. Everyone is not privileged to get university education. People taking online courses may not be motivated all the time. Somebody has to direct people to organized the online materials. One should decide how much "AI" is necessary for them while searching for the content.

We need to know about the intent of people getting into the field of AI. Virtual reality in education could be explored. An important aspect of this is to understand – How much mathematics(probability, programming etc.) is required. What is the necessary background to be proficient in AI? There is a need to raise the awareness regarding the hype about AI. Such introspections brings us to more concrete steps in terms of what is required from the policy of AI education in school.

Policy of AI education in school

- How to implement AI education in school?
- How to educate educators?
- What are you going to replace in terms of courses? Or do we really need to replace any course in order to place AI in school?
- Making sure that the courses are uniform across the country or region.
- Recent steps taken by most of the states in US to make mandatory exposure for programming or data analysis.
- How exposure of AI in high school can help? There are some studies show that when students are positively introduced with any subjects in middle school, it will lead them to pursue the same field.

1.1.4 Autonomous Vehicles

Session Moderator: **Hager Radi**, American University in Cairo
Session Note-taker: **Khimya Khetarpal**, Montreal Institute of Learning Algorithms

There were 5 attendees in this session. Key topics discussed were levels of Levels of autonomy in vehicles Level1-Level5, definitions of these different levels,

Role of autonomous vehicles, Do we need this technology? The role of autonomous vehicles depends on the level of the autonomy. This is be a valuable technology but not necessary. It offers accessibility to elders, blind, It is very useful to serve for human error, thus offers huge benefits in terms of safety and productivity. It requires systems for understanding of the task to ensure safety for different categories such as sensors, perception, planning and control. In fact, without limiting to autonomous car, possibilities are endless: such as auto drones, delivery of packages in scenarios where drones helping basic health packages to certain areas humans cannot access for e.g. nuclear sites investigation. Others might include pizza delivery by AI.

Adverse effects of this technology? Automation leads to job loss. However new jobs that will come by would be much better and less labour cumbersome Jobs loss is arguable to a small extent as even today: more demand than supply of drivers, so automating this will help us and not harm us. Will this not replace humans entirely? Human in loop in high density areas like downtown.

What technology is used in cars today: Are cars using DL based models? Tesla may be is using neural nets based models. One of the participants mentioned that, 3D object based detection is deployed on car but such system requires high levels of optimization and it is very difficult to standardize. Training and testing can be done using a Titan GPU, but it can be never used in actual cars. There are heavy constraints in computation and energy consumption. Inference is much less compute expensive, so cloud could be used for training and more data computing. Supervised learning offline is the trend and most used so far. More examples of cars using few shot learning and self supervised training are yet to be seen.

Would this be a common man's dream or only the chosen few? Today it is only a luxury to have level 3 features such as distracted driver detection, lane detection, park assist etc. Open areas include research needed for optimization and cost cutting to make this technology accessible by the masses

Ethics alignment in Self Driving Cars Value alignment is still an open question when it comes to autonomous driving vehicles. MIT is collecting data on who should a machine kill- an old lady, a baby on the road? Humans also do not know this, how would the machines know? A tricky question is if you put a machine in this situation. A common worry of researchers is media hypes etc.

1.1.5 Time Series Modeling/Forecasting

Session Moderator: **Patrick Kakou**, Ryerson University
Session Note-taker: **Odd Sandbekkhaug**, Infiniwell.ai

About 40 attendees joined this session and session was about discussion on Time Series modeling in medical applications, and the difficulties therein particularly considering the complexity of problem space and the difficulty in definitively labeling complete data sets.

Common problems with using time series data: Patients only come to doctor after disease sets in, and so most data sets only reflect an already sick population. Data sets are biased by the person doing the labeling (interpretation). There is a need to standardize terminology and try to remove the human bias from the data set labeling. As a result, models do not perform well on alien data sets. Models can behave well on a given population data set but fall apart when data sets are exchanged. Note: In a subsequent session it was noted that even medical imaging has this problem, where the data set is prone to be biased by the image scanner and processing techniques. A model trained on images scanned at one hospital may not perform well on images from another scanner/hospital.

Medical conditions can be difficult to diagnose (or label in other words), and doctors may not always agree on a diagnosis when looking at the same data. We must distinguish between syndromic conditions (multiple possible causes) and conditions with a clearly defined physiological underpinning. Syndromic conditions are more difficult to model; the latter may be more tractable for time series modeling. Example of conditions that are not syndromic - diabetes: can measure blood sugar level and consequences of blood sugarâĂę impact on liver, retina etc. That is a tight condition. Good yet simpler models can be developed around very tight conditions, but finding a generalized model is a long way away. Complex models do not generalize from population to population

Data analysis perspectives from other domains:

- Netflix for instance has more well-defined problem spaceâĂę no ambiguity of target which makes modeling easier. Demand modeling is at the aggregate (population) level and ignores variations in individuals.
- Climate modeling: If the outcome canâĂŹt be clearly modeled, then try to incorporate knowledge of physics models to improve prediction models, thereby embedding a level of expert knowledge into the model
- Aerospace: work with domain experts to identify the key parameters, then start modeling from that as a starting point rather than trying to develop a general model based on a wide data set

AI models can be used to educate practicing physicians rather than using AI for decision-making. Example: Massachusetts general running AI models on radiology images before showing images to doctors; improves overall radiologist performance. Potential problem with time series modeling and prediction on a daily basis: if the model has some noticeable volatility, it can erode stakeholder trust even if the model is accurate. Too much volatility in prediction is bad even with high accuracy.

The Complex Systems Problem

Biological systems are complex systems, and therefore by nature difficult to predict. For time series models this is especially difficult as complexity increases [exponentially] with time and number of parameters. The most successful ML applications are reported on static data sets with no time context such as image processing. Syndromic conditions are difficult to predict because they may have many different contributing factors. As one example, doctors do not agree on a single clear definition of the causes of sepsis, and so it is difficult to develop a generally effective AI model. Similarly, seizures can have many different causes with no clear physiological underpinning and is therefore difficult to model/predict. Generalizable algorithms that work on exchangeable data sets is the holy grail. Difficult to achieve for health-care, easier in physical systems.

The Data Set Problem

As discussed before, clinical data sets are inconsistent, incomplete, are weighted towards people that are already sick and reflect the bias of the person doing the labeling (interpretation). Models trained on one data set do not easily generalize to other data sets. Models trained on one data set perform poorly on another data set, so there is a definite problem of inconsistency here. There is in addition the inherent person-to-person variation which makes constructing a generalized model difficult. More work is needed to standardize terminology and remove the human bias. If data requires human processing, then the data set is biased towards the human thinking rather than the disease process. Remove the human thought process and get better raw data.

Available health care data sets are typically weighted towards sick people, as data is usually collected only after a person has become sick. We do not have similar/comparable data sets on healthy individuals. One possible way to offset this

bias: supplement data sets with bio-metric data from healthy individuals through data from Fitbits, smartwatches etc. from healthy people. However this has its own set of problems: people with smartwatches may come from similar socioeconomic backgrounds and have similar [better] health profiles. They may not reflect the general population but rather a smaller segment.

Traditional diagnostic processes use decision-tree models (i.e. an algorithmic approach) today with good success. However decision trees donâĂŹt work well for biological systems due to their complex nature where there is not always a clear cause-and-effect. AIâĂŹs challenge is to improve on that. Predicting time series data in biological systems gets exponentially difficult with more parameters and longer timescales. For AI, we might instead consider picking non-syndromic conditions with a clear physiological underpinning: narrow the scope of the problem by picking a disease process that is as âĂIJtidyâĂİ (or more deterministic) as possible and develop AI models that perform well in this space. Better to perform well on a narrow problem than perform poorly on a wide problem

Takeaways

There is a need for better and standardized health-care data sets which are less biased by the humans that label them. For biological [complex] systems we should narrow the problem space and aim to solve simpler problems first. That is, find a well-defined subspace that is amenable to time series modeling.

1.1.6 Disentangling Session

Session Moderator and Note-taker **Ben Poole**, Google Brain

There were about 30 participants to this session. This was one of the sessions proposed during Day 1 and self organized the same day with a huge turnout.

What is disentangling?

There were several definitions and understanding participants have for disentangling. In a nutshell some sort of clear separation. But is it really? These included many different versions enlisted below.

Various definitions of disentangling?

- Feature wise control for RL
- Representation learning: e.g. style, content
- Causality: disentangling as interventions, separating clusters spatially within a latent in space,
- Sparse distributed representations, discriminating ability even if not close such as each feature represented in different axis <-> de-correlation
- Individual components are all separable
- Features being orthogonal to each other. Does the basis matter? Why axis-aligned?
- Uncovering latent factors in the data-generating process
- Independent and interpretable
- Useful for causal/intervention studies
- Recovers the true data-generating process
- Discovering laws of physical world

Properties of a disentangled representation

- Independent and interpret-able
- Aggregating posterior independent
- Interpret-able to humans

We can check with generative models that the components are interpretable images capturing particular generative factors e.g. hair color, smile, etc.

Measuring disentangling? Simple linear readout from latent space -> task you care about, e.g. binary attribute classification. Can we define disentanglement in the absence of downstream tasks? In fact Why focus on VAEs?, Do GANs also disentangle? People donâĂŹt think about GANs in terms of that but we donâĂŹt know whether they disentangle (needs more experiments) How about the prior space already independent in terms of dimensions?

Is this restricted to continuous processes? What about discrete processes? E.g. sequence of spoken works is a sequence of discrete words. How much of âĂIJdisentanglingâĂİ comes from constraints on prior? What about other kinds of distributions. If latent is Gaussian, are there better âĂIJdisentangledâĂİ representations in the intermediate hidden layers of the decoder network?

Related work might include intrinsic image decomposition Janner et al. [2017] and natural decomposition of images into shape, reflecting, etc. that are inherent to images. Some thought one could encode additional inductive biases into CNNs, e.g. group symmetries, factorization, etc. Another aspect is disentangling in terms of recovering the true data-generating process. Is this magic when it works? Relaxing away from recovering true generative process. If we had true generative process, what would we do with it. Perturb latent space, see how that impacts data distribution? How would you check that you learned something useful about the world?

For many generative processes, there are underlying physical laws, why canâĂŹt we discover these? A few or one of the participants even proposed that — we stop using the word disentangling. An interesting question that came up is — Are there disentangled representations in human brains? Disentangling as a verb vs. a state of something. Model extracts something true about reality vs. disentangled state âĂIJcan use itâĂİ. Tautology: are we just saying

that disentangled == good? Is there such a thing as too disentangled? Attributes can be correlated, e.g. beard and man, open/closed mouth and smile, donâĂŹt want latent space that are totally independent.

Discussions included neural vs. symbolic disentangling, localized vs. distributed representation. How much sparsity do you need? Do disentangled representations have lower capacity ? This may have nothing to do with interpretability. We do not understand the world but we are still operating in the world

Why Disentangling

Some people think that this is needed for generalization wherein one could holdout some subset of states for factors. *Pragmatist disentangling agenda vs. fanatic view on disentangling:* Pragmatist outlook says independence gives us more control over the system where as the fanatic view says reveals something about the universe.

Disentangling should help with explainability: Find factors that are driving the system, explanation. However we need some framework to describe success in unsupervised learning. All the ways we have described it could be some goal. Disentangled: Can train a classifier vs. tweak variables and see impact of simple transformations. Is it even possible? Given a data-set can we know that is disentangled? Could you get the best possible disentangled representation? How do you measure this? Not just one solution? There may be many possible solutions.

PCA vs. ICA

Disentangled features may improve continual learning. At the same time what about supervised learning? Where is disentangling in Imagenet? Metrics for disentangling could include linear readout of variables you care about which would depend on the data-set. In the case of supervised learning as disentangling: a classifier can be used to infer one disentangled latent variable (e.g. class). In context of feature wise control/continual learning, evaluation/metrics could be on synthetic tasks where we have access to data-generating process But may be too simplified, may not represent progress.

Why is disentangling not tautological in practice? ItâĂŹs a hard unsolved problem on toy tasks where the true underlying factors of variation are factorized. Disentangling is how well representation generalizes to new tasks, but then Who gives you that new task? May be hold out particular combinations, e.g. leave out certain positions/combinations Zamir et al. [2018]. More realistic benchmarks between toy and real such as video game engines do know the factors of variation, donâĂŹt map onto generative models and x-ray traced scenes are sufficiently complex, interesting 3d geometry.

Caution on interpretability: Often ascribing certain things to representations. Easy to fool humans and easy to create post-hoc justifications of what knobs mean. Some instances are meaningful, other instances hard to know what itâĂŹs controlling. How do we know that we have been successful in new domains, e.g. text/biology/etc.

Disentangling as inductive biases: Discovering true features that are relevant for the task When is disentangling useful?— Human-in-the-loop generation, Feature control, want the knobs to be meaningful to that system, independent optimization over different variables, makes downstream optimization tasks easy.

Pessimism: Maybe VAEs are not working as well as we think they are in the image domain. Language is already a bottleneck, why/how to force it through an even smaller bottleneck? Language may be compressed, but not in a format that makes it easy to do other tasks (e.g. sentiment classification). In, CV tune knob can change features on human face: Post-hoc justification of these knobs. Elephant in the room: presumption that VAEs work. There are optimization issues Become compounded as you move to other domains. For Natural Language, we want control that CV has for generating text. Why are not we there yet? Is it because of the discrete nature of text? Range of values much smaller, e.g. grammatical correctness, previously generated tokens, canâĂŹt repeat, etc. Embedding for each token, tried perturbing embeddings.

Is disentangling a natural consequence? – If we just train w/more data/longer do you naturally get something like this. Compression is not enough: If compress down to entropy of data, very efficient representation but not interpretable. Animation generation, in between data-sets between synthetic and images What does synthetic mean in terms of text? What should be the controlled factors for text generation? Additional reading: Weston et al. [2015], Sabour et al. [2017], Devin et al. [2018], Achille et al. [2018], Whitney [2016]

1.2 Day 1 Session 2

Session 2 took part between 1-2 pm. The session topics were *Reducing Dependency on Massive Labeled Data-sets*, *Language Modeling and Machine Translation*, *Justifiability, Explainability, Interpretability and Algorithmic Decisions*, *ML for Creativity*, and *AI Governance/National AI strategies*.

1.2.1 Reducing Dependency on Massive Labeled Data-sets

Session Moderator: **Collin Raffel**, Google Brain
Session Note-taker: **Jorn-Henrik Jacobsen**, Vector Institute and Trevor McKee, A.I.Vali Inc., Starr Innovation Centre

In this session there were 38 attendees. The session was divided into multiple subgroups so more specialized topics could be discussed. One group discussed general approaches to small data learning, semi-supervised learning and transfer learning. Another group focused on medical imaging data and how to incorporate domain knowledge into machine learning systems.

Group: Small Data, SSL and Beyond: First we discussed real-world experience with small data. Some mentioned examples were:

- Auto encoder on molecules, latent space as unsupervised representation.
- Reinforcement learning with sparse rewards
- Mixed modalities in heterogeneous data-sets (e.g. medical), you can always do decision trees / random forests, but does not scale to many features
- Leveraging additional information like hierarchical relationships between labels

Next discussion was on – why is there no breakthrough in unsupervised learning yet, or did we miss it?

Supporting Key Points:

- Self-supervised pre-training in NLP! BERT is biggest unsupervised success story. But is this really unsupervised? Someone argues rather pre-training with cheaply obtained labels.
- Transfer learning: Transfer learning can give clear gain, even when there is no overlap between train and target data sets.
- Real-world examples of transfer learning: Instance-based transfer learning can give win.
- Weak labeling: CV and NLP communities use it.
- Intrinsic dimensionality of data-set. Fix the subspace dimensionality.
- n+1 prediction. It might work for some high level representation, but n+1 prediction does not seem for sequences of words. However, VQ-VAE does discover phonemes from raw audio wave-forms.

Controversial Key Points

- No breakthrough yet in unsupervised learning.
- What is success? Generative models need to be contextualized to be evaluated
- GANs: Why should they be useful? Maybe same issue as any generative model. Maybe not trustworthy, as we do not know what we are optimizing.
- How about tasks that are not useful? Adversarial training mostly hurts test accuracy. VAT on the other hand helps.
- Multi-task learning, how to infer if new tasks are useful? Taskonomy is a paper in this direction Zamir et al. [2018]. Maybe things are great if we are more careful in our definition of tasks and domains.
- How to choose subset of data for unsupervised or supervised pre-training. Core sets might be one approach?

Group: Medical Image Data / How to Incorporate Domain Knowledge:

One âĂIJconclusionâĂİ that we came to was that the more specific or constrained the question is, the less data you might need - an example: you either have two buckets of imaging data labeled âĂIJsickâĂİ vs âĂIJhealthyâĂİ, versus data-sets with segmented regions showing the exact location of the lesion: the latter will likely require less data to produce a robust classifier. Reference to anatomical atlases, or public data-sets to train against, can also help (with

respect to rare diseases, where it is potentially just not possible to obtain enough data for training - learning latent variables that somehow describe the system can help to identify salient features identifying the rare diseaseâĂŹs difference in genes or image features, perhaps).

Takeaways

Small-data efficiency can be achieved in various settings. The problem is that we often care about very specific properties in our representations to solve downstream tasks well. Therefore self-supervised pre-training with cleverly designed auxiliary tasks and objective functions closer to the domain of interest often work better than generic unsupervised pre-training approaches. Open questions are if this is a fundamental consequence of the definition of unsupervised learning, or if all existing approaches can be cast into generic frameworks, so the amount of feature engineering in e.g. self-supervised settings can be reduced.

1.2.2 Justifiability, Explainability, Interpretability and Algorithmic Decisions

Session Moderator: **Gillian Hadfield**, University of Toronto
Session Note-taker: **David Madras**, University of Toronto

Number of Attendees: 30

ModeratorâĂŹs Presentation What are people thinking about when they think about explainable AI?

- Why did my model succeed/fail? When to trust the model?

What about interpretable AI?

- WhatâĂŹs happening inside the model? E.g. Chris Olah et alâĂŹs work on feature visualization Olah

Both of these (model success and interpretability) provide information for builders and users.

Are there other types of explanations we need?

For example, the GDPR: we may require other explanation types.

- The GDPR may (or may not) require an âĂIJexplanation of the decision reached after such assessment and to challenge the decisionâĂİ
- Is this the same type of explanation as above?
- Legally, we want human accountability and justification.
- Giving reasons for decisions, consistent with what a community judges as acceptable.

How does our society depend on institutions to classify?

Can we connect algorithmic decisions to humans to hold them accountable?

John Rawls: to believe in our legal system, we need good reasons for the judgments that are made.

- Supreme Court of Canada: âĂIJThe delivery of reasons is inherent in a judgeâĂŹs role.âĂİ

Can we develop a procedure for licensing algorithms to make sure that algorithmic decisions can be justified?

- Maybe train a human to âĂIJlicenceâĂİ a model?
- Difference between explanation/justification/interpretation must be defined.

Discussion

- Justification: does this only make sense for decisions that humans can articulate reasons for? What about object recognition?
  - ⋆ The domain of justification is exercising judgment.
  - ⋆ The domain of explanation is wide.
- Some judgments require justification - norms and institutions matter.
- Do saliency maps constitute an explanation for a CNN?
  - ⋆ Making pretty saliency maps that align with our intuition is not usually a simple process.
  - ⋆ Just using SGD does not give you anything too interpretable on its own.
- Two approaches to interpretability: we can either look at what happens when we change a feature, or we can build a property directly into the model.
  - ⋆ But low dimensional explanations are pretty unsatisfying.
- What comes after feature attribution?
- Maybe we can use expert knowledge to develop better abstractions.
- How good can our justifications be if we also develop them for bad ideas?
  - ⋆ In a legal context, they must be âĂIJsubject to critique and reviewâĂİ âĂę we need norms.

- How can we reason about the behaviour of AI systems?
  - ⋆ Natural language explanations?
  - ⋆ Work with RNN and attention - a mask pointing at different parts of data for each word.
  - ⋆ Train a model to output a post-hoc explanation?
- Maybe the only real use of explanations are for model debugging? :(
- Causality seems relevantâĂę
  - ⋆ If you have structure and prior knowledge, you can use that to build âĂIJexplainabilityâĂİ into the system.
  - ⋆ Outputting counterfactuals is one method of interpretability.
- What is an explanation anyways? There are different theories of scientific explanation.
  - ⋆ Maybe an explanation is an argument.
  - ⋆ There are distinct use cases for explanations: causal reasoning, creating an argument, etc.
  - ⋆ Explainable AI: maybe there are insights from the social sciences.
  - ⋆ These types of explanations are important if we want to learn more about the model, but maybe not if weâĂŹre just debugging our model.
- The purpose of the explanation matters.
- What are good metrics?
- Maybe there are other interesting domains, where there the relevant structure is global, rather than local.
  - ⋆ For instance, thereâĂŹs no saliency map that explains why an image is blurry, or why a word is a palindrome?
- Is justification only post-hoc? If it were proactive, would that be like transparency?
  - ⋆ The fact that we know we will have to make a post-hoc justification affects our actions anyways. Norms are related here too.

Takeaways

- Feature attributions/local explanations are somewhat unsatisfying, what else can we do?
- We can only really evaluate an explanation well if we know what it will be used for
- Explainability, interpretability, and justifiability are all different, and all important.

1.2.3 ML for Creativity

Session Moderator: **Tom White**, University of Wellington
Session Note-taker: **David Kale**, Netflix

Number of Attendees: 13

Two broad classes of problems: Building a tool or system that is inherently creative and Building a tool or system to assist humans in creative endeavors. The challenge is evaluation: how do you evaluate the quality of generated or created content?

Style transfer: no immediate business application but broad uptake, enabled new kinds of thinking about how to apply ML. Why was style transfer so successful? Was it by chance? Is it easy to use or well-suited to Internet meme culture. Why more successful than deep dream? — some of the reasons brainstormed were easier to explain, more obviously useful, response from creators: forgery?. Novel but not too much (Familiar enough to be approachable)

Is creativity an inherently human act? Is creativity more than sampling from a semi-random process or interesting latent space? But âĂIJtasteâĂİ may be creative act: people could sample from, e.g., GAN, and retain samples they like – thatâĂŹs creative! Also, creativity can be a property of an overall system, a la Netflix (which distills member interests and then hires human creators to make content aligned to those tastes)

Should we try to define creativity? Rabbit hole: might get bogged down in debate while never settling on a potential solution. Creativity could be Latent space interpolation (not a training data, further away in the latent space than the training data). What defines art ? — Person in the loop making a choice, or taste (i.e. filtering / rejection sampling based on prior) and photography led to the creation of modern abstract art. Generative arts could be the next step. ChristieâĂŹs auction stirred up a lot of interest, journalists used it to stoke fear.

Authorship Can an algorithm be considered an artist, or is it a derivation (forgery?) of its training data? How can we remove the âĂIJbiasâĂİ of the source author? Maybe authorship will be separated into groups (data, developers, researchers, artists) just like movies have producers, writers, actors, etc.

1.3 Day 1 Session 3

Session 3 took part between 2:45 pm to 3:45 pm. The session topics were *AI for Social Good*,*Interesting, Hard to Explain Phenomena*, *ML in Supply Chain*, *Curiosity-Driven RL* and *Technical AI Safety*.

1.3.1 AI for Social Good

Session Moderator: **Bogdana Rakova**, Samsung Research America
Session note-taker: **Nikolaos Sarafianos**, Universty of Houston

Brief information about the session:

UN has metrics for Social Good: 17 development goals, difficult to measure, 150 indicators to measure progress to measure progress of the 17 goals. How the ML community can support existing indicators. The UN General Assembly has developed a global indicator framework which contains 232 unique indicators, addressing each and every one of the Goals and targets of the 2030 Agenda for Sustainable Development.

Key takeaways from the latest UN SDGs report:

Quality data are vital for governments, international organizations, civil society, the private sector and the general public to make informed decisions and to ensure an accurate review of the implementation of the 2030 Agenda. AI tools are not equally distributed - UN is working on a road map for the modernization and strengthening of statistical systems.

We need to consider about the interlinked nature of the SDGs. ItâĂŹs ultimately very difficult to measure the complex ways in which they depend on each other and therefore itâĂŹs crucial to have interdisciplinary partnerships between civil society, academia, the private and public sector (to inform the production of more evaluation indicators).

Key Topics Presented:

- UN sustainable development goals
- Quality data and how unevenly distributed the tools are.
- ML to Health care (growth patterns)
- Privacy vs good quality data. How do we think about data and privacy.
- Lack of talent in other regions of the world can impact the possibilities of solving such problems. The cost of technology is very expensive in other parts of the world. Besides ML you need to have other stakeholders in the conversation (governments).

Perhaps we want to first build community then try to convince other stakeholders about the impact of such technologies. Besides the session discussed: What is the concrete of problem that we should focus for positive impact. Governments are averse to risk. As a researcher you have risk attached to you. The problem might not be technology but politics/advocacy so as to convince key players. Check AI superpowers book. We need more companies to join the conversation.

How to balance âĂIJmaking moneyâĂİ vs âĂIJsocial goodâĂİ. All the presence of companies in Africa was only on business case. Data that comes from the public is then going to the company directly. How to get funding to do good. How to deal with data collection so as users will trust you that they will not be used. Who structures it in a way so as people are comfortable. Provide educational resources, unrestricted funding. Do outreach to bring more people who will then go back and fix things. Hackademia campaign in different places to teach people electronics hackadamia. Totally different applications in different parts of the world.

A company can make the data accessible to the research community. If information is available to everybody then Data can generate value for the owners who uploads it. Selection bias can be introduced. If the incentive is money you get different data than general population data. Given different backgrounds if you put money as the incentive to get data then you have a skewed representation. A person in deep poverty will find the financial incentive attractive. Maybe data can be made anonymous. Having good use cases could be worth sharing (how we set things up). Start with simpler examples and build them up over time.

Maybe we need to do AI 101 for governments to make them get into conversations. If you donâĂŹt do that you might look like a magician to other stakeholders. How to empower people with technology and resources to build their own things in other parts of the world. Pairing academia and communities around the world could be a solution.

Practical examples include:

- Predicting weather 7 days in advance: In Tanzania for small-scale farmers thereâĂŹs no data connection. Can I build a predictive model when I have access once a week to the internet? If this work improves the farm planning.
- Build disease detectors for leaves: Smart phone takes photo of leaves to find diseases. Challenge is find the places that itâĂŹs appropriate to use such tools.

Database of concrete problems -> Matching system between people with ideas and people with problems to find good solutions. (UN might have something like that) Collectively aggregating different data-sets could be useful. People should be able to provide their data if they want to? Privacy by design? Summit

1.3.2 ML in Supply Chain

Session Moderator and Note-taker **Miti Modi**, University of Toronto

Brief information about the session: There were 7 attendees in this session. McKinsey and Company recognized supply chain as one of the two applications of AI with the largest potential for economic value hereInstitute [2018]. Supply chains include planning, sourcing, procurement and management of logistics. These networks can be very complex and can span globally here of Supply Chain Management Professionals. In this session we discussed some of the existing challenges in adoption of ML in supply chain and future directions.

Main discussion points:

Forecasting: ML can be applied to solve several challenges within forecasting demand. Some examples are:

- Use product + store features to forecast demand for new items (ex. Fashion changes fast, new items introduced to customers very often and it is a common problem for retailers to accurately forecast demand for the new products)
- Introduce sizing as a product feature in forecasting for fashion as currently forecasts are generated at an item level and not at an item-size level
- Predict Demand Transference ie. How does the demand of product that is out of stock transfer to other products on shelf.

Online vs. Physical stores Application of ML in supply chain can vary depending on whether the store is online or has a physical location:

- Physical stores rely on point of sale (POS) and basket level data (ex. What brands were sold together). POS data is a very small percentage of what actually happens at the store vs. being online is an opportunity as you have data available about your customerâĂŹs movements, about the products they were browsing but did not end up buying.
- In physical stores you actually have to send the products to the stores, vs. in online stores you have more opportunities to experiment with new products

Price Elasticity: Price elasticity is a measure of how changes in price of products impact their demandhere Academy. ML can be applied to determine optimal pricing for products that can maintain customer's satisfaction and also improve retailers' profitability. However, applying ML to determine price elasticity would require A/B testing. Therefore, this would be easier to do this online vs. in physical stores. Currently, most physical stores rely on printed price labels and so they would need to invest in electronic shelf labels to carry out A/B testing.

Bias in Data Retail data can be biased. One example is that humans select the products that will be sold at stores. If these products sell well, then this can confirm the merchandiser's bias related to the types of items that the consumers need. However, we have no data about the products that were not picked by the merchandisers. What if a product that was not selected by the merchandiser could have also sold well? Most supply chains are used to being heavily reliant on humans.

Future directions/Takeaways: Some future directions for adoption of ML to supply chains are:

- It will be important to convert items, brands etc. to word embeddings (like word2vec models) for forecasting.This can help forecasting for new products.
- Supply chains will need to start fostering a strong culture for supporting innovation.
- Physical stores will need to make investments (ex. electronic shelf labels) and will have to be creative in their application of ML to compete with online stores.

1.3.3 Curiosity-Driven RL

Session Moderator: **Feryal Behbahani**, Latent Logic and **Khimya Khetarpal**, Montreal Institute of Learning Algorithms
Session Note-taker: **Erin Grant**, University of California, Berkeley

The session was started by a brief intro to the topic and literature followed by the discussions. The session moderators put together a guiding document for the discussion which they plan on publishing on a blog soon.

Key topics discussed were:

- Where do rewards come from? What signals promote learning?
- Why curiosity matters in RL? Background on the origins of curiosity for learning.
- Types of curiosity signals proposed in literature.
- Recent success stories of curiosity-based learning agents.
- Identify open problems and closing discussions.

The *prime focus of this session was* — What makes an agent act?! What is the motivation of the agent? What kind of reward functions should we use for learning? There are forces in the environment that give reasons for the agent to direct its activity, these can be classified into 2 broad groups namely extrinsic motivation i.e. acting to maximize some external reward from the environment (e.g. get a higher score on Atari) and intrinsic motivation — do something just because it is interesting and inherently satisfying or enjoyable, for its own sake (e.g. play, explore the environment, or even learning itself!)

Curiosity-driven learning and intrinsic motivation have been argued to be fundamental ingredients for efficient learning Freeman et al. [2014] What are the states of curiosity? What happens if curiosity is not well directed — it can cause bad behaviour which keeps the agent from exploring useful part of the state-space. This led to the discussion Noisy TV problem.

Discussion included kinds of intrinsic motivation/curiosity and formulations including 1) Measuring Novelty: explore unseen states Bellemare et al. [2016], Grondman et al. [2012], Poupart et al. [2006] Prediction Error: perform actions that reduce the error/uncertainty in the agentâĂŹs ability to predict the consequence of its own actions. Discussions included around what is drive for humans. How do we act ? Is it really curiosity when we do a specific task or it is a bigger task framework called life. Participants discussed ideas from psychology literature.

Related work has modeled curiosity/boredom to improve a world models. Besides, reward has been based on changes in the prediction error not directly on the magnitude of error. More recent research on World models Ha and Schmidhuber [2018] is also in this direction. Also relevant to his later theory of Compression: Driven by Compression Progress.

Discussion revolved around understanding how curiosity has been used in RL to explore. However there are open questions as to if this is sufficient and all we need curiosity for drawing parallels to human reward model and day to day lives. Additional Reading: Schmidhuber [1991b,c,a], Itti and Baldi [2009], Berlyne [1960], Friston et al. [2006], Oudeyer et al. [2016], Burda et al. [2018]

Open Questions:

- What is the best environmental setup for studying the effects of curiosity?
- Is rewarding exploration through curiosity always a good idea?
- Is there such a thing as safe curiosity signal?
- What should the agent optimize considering there are intrinsic and extrinsic rewards to be considered?
- Teacher student learning (i.e. automated curriculum learning) would be also worth discussing and how these measures of curiosity can be leveraged there.

1.3.4 Technical AI Safety

Session Moderator: **Victoria Krakovna**, DeepMind and **David Krueger**, Montreal Institute of Learning Algorithms
Session Note-taker: **Tegan Maharaj**, Montreal Institute of Learning Algorithms
Number of Attendees: 18

Brief information about the session: Overview of the field of AI safety, then discussion about technical advances

Key Topics Presented:

- What is AI safety?
  - ⋆ Near term AI safety:
    - ☐ Issues we are facing or will face in the next decade.
- Long term AI safety:
  - ⋆ Issues we may face with more advanced AI decades from now.
  - ⋆ Session focuses on these type of issues, as well as connections with near-term problems
  - ⋆ **Main Topics:**
    - ☐ *Specification:* define the purpose of the system while reliably specifying human preferences.
      - ◇ *Value learning:* How to reliably specify? (GoodhartâĂŹs law makes this hard - when a metric becomes a target it ceases to become a good metric)Chrystal et al. [2003]
      - ◇ Value selection: whose preferences, how to aggregate?
    - ☐ Designing AI systems to act in accordance with these preferences.
    - ☐ *Specification gaming:* Krakovna
      - ◇ Pause tetris indefinitely to avoid losing.
      - ◇ Win race by being really tall and falling over fast.
    - ☐ *Robustness:* Indifferent to perturbations:
      - ◇ Distributional shift adaptation, or at least fail gracefully.
    - ☐ *Assurance:* check that what you think is happening is
      - ◇ Interruptability (indifference to shutdown).
  - ⋆ **Technical approaches to these problems:**
    - ☐ *Interpretability / Adversarial examples*
    - ☐ Safe Interruptability
    - ☐ Avoiding side effects
    - ☐ Safe exploration
    - ☐ Learning reward functions
      - ◇ From preferences
      - ◇ From demonstrations (inverse RL)
      - ◇ Specify / encode instructions / demonstrations (i.e. meta learning)
      - ◇ GoodhartâĂŹs law is still a problem (experimental demonstration that agents can learn to game the learned reward function - not just engineered specifications are vulnerable; even learned specifications can be gamed)
      - ◇ Instrumental goals can also be a problem.
      - ◇ Imitation learning (behavioural cloning / teacher-forcing, often similar to reward learning), less ambitions because more is specified about how the algorithm should behave. Question Answering system (oracle) e.g. Amplification, explanation by debate.
    - ☐ Improving theoretical foundations.
      - ◇ Do we have the âĂIJrightâĂİ way to make good AI?
      - ◇ Assurance for current systems (interpretability, adversarial ex., verification)
      - ◇ General RL theory e.g AIXI (more general than RL so may scale better).
      - ◇ Beyond RL (âĂIJembedded agentsâĂİ) - flow chart of problems in slides.
- Are there exceptions to GoodhartâĂŹs law?Can we try to achieve these exceptions ?
  - ⋆ e.g. temperature and homeostatic regulation (optimizing for temperature).

  - ⋆ This depends on size of search space; can still apply.
  - ⋆ Temperature is a well-specified quantity (causal models are well understood so we can know that it is causally predictive of the quantity of interest; causal model related to size of search space).
- How can we be confident that AI systems will operate as we expect or intend?
- Possibly can solve a lot of these problems via better learning from sparse rewards.
  - ⋆ Less vulnerable to overfitting / GoodhartâĂŹs law. However there is still a specification problem.
  - ⋆ Vulnerable to the values of the sparse rewards, but may give humans more time to give better rewards (but humans are bad at giving these rewards generally).
- How to model reward breakdown (e.g. âĂIJI want to go to collegeâĂİ, intermediate reward for going on reddit).
  - ⋆ Few solutions might include using different scaling factors, using different discount factors (analogous to prioritization in humans).
- Multi-objective optimization.
  - ⋆ Try to give yourself more options for trade offs between objectives.
- Commonalities between near-term and long-term objectives:
  - ⋆ Short term: make decisions fairly, Long term: learn from human preferences (both of these have hard-to-pin-down concepts that are shared).
  - ⋆ Better causal models helpful for both near and long term.
  - ⋆ If we canâĂŹt solve the short-term the longer-term will be harder.
  - ⋆ Adversarial examples: What weâĂŹre dealing with currently is not the problem we will actually face; how to do we scale to relativity? this scaling is more general than just this problem.
  - ⋆ Fake/generated things (telling what is real),
- Differences near vs.long term.
  - ⋆ How much can we tolerate failure? In many cases, we are okay with some small possibility of being bad / tolerating some risk, but there are cases where if advanced AI made just one mistake it could be catastrophic and we do not want to tolerate that.
  - ⋆ Telling if something is âĂIJfundamentalâĂİ or less fundamental; how well things will scale e.g. coming up with metrics of fairness is independent of the model used for classification etc.
  - ⋆ What does âĂIJscaleâĂİ mean?
    - ◇ Must be applicable.
    - ◇ Must be influential.
- What determines the difference between near and long term safety problems?
  - ⋆ Scale of AI, both in âĂIJamountâĂİ of intelligence/how good it is, and also in how broadly they are deployed.
- Why would we ever put an AGI in a position of control?
  - ⋆ Does not have to be fully AGI in order to start causing problems for instance failures in airplane automatic flight system due to faulty sensor.
  - ⋆ Several ways this could happen is an arms race in military, finance, technology, smaller powers wanting to have more power.
  - ⋆ need to represent populace / do social organization.
    - ☐ Currently markets, democracy, social choice theory include putting new agents into the model changes the equilibrium; preferences are relative / donâĂŹt exist.
    - ☐ Instrumental goals (in theory, all optimization algorithms will instrumentally try to get certain things including power and resources).
    - ☐ Indirect control (say something on Facebook that causes someone to do something in the real world; even though the algorithm only had access to a text box it can affect the world in ways we donâĂŹt understand).

- ◇ Expert manipulators such as accidents (e.g. flash crash)Sagan [1993].

- At the same time, the advantages for AGI include being faster, being more âĂIJrationalâĂİ (perfectly optimizing). However some believe that this is not actually always an advantage; being rational makes you predictable and in game theoretic / economics can be less important.

Open Questions:

- How can we tell which approaches are likely to scale?
- What alternatives are there to reward maximization / optimization
- What are the low-hanging fruit for safety? (making weapons not connected to the internet or to each other, keeping humans in the loop for safety)

2 DAY 2

There were 15 discussion sessions planned for Day 2. They were structured similarly to the sessions from Day 1.

2.1 Day 2 Session 1

Session 1 as the first session of the day took part between 10-11 am. The session topics were: *Neuroscience-inspired ML*, *ML for Climate*, *Diversity and Inclusion*, and *Inverse RL.*

2.1.1 Neuroscience-inspired ML

Session Moderator: **Simon Kornblith**, Google
Session Note-taker: **Meltem Atay**, Middle East Technical University
Number of Attendees: 16

Brief information about the session:

In this session we discussed about why we need neuroscience inspiration and how is it possible to make ML methodologies better with such inspiration? During the session our main question was âĂIJWhy we need neuroscience inspired ML?âĂİ We mainly focused on obstacles of understanding brain thoroughly and several obstacles on applications. Due to advanced structure of the topic 3-4 people dominated all the conversation some left early.

-Supporting Key Points:

- Human brain ability to generalize with low data and low energy should be deeply understood to be mimicked for ML systems.
- Strong representation and transfer learning of humans should be more efficiently applied to ML
- Learning from weakly labels and transfer learned or inherited representations is important
- These applications would improve the generalization of RL
- Understanding and explaining visual system was so comprehensive so ML approach was able to mimic this system using its own rules.
- Getting connectivity matrices ,modelling different states of the brain would provide more valuable insights.

-Controversial Key Points

- Considering applications of RL, cognitive science is providing more insights than neuroscience.
- Mimicking human brain requires complex language representations since human learning relies very strongly on language, comparing to other mammals.
- Some discussion occurred about how to apply this idea on ML or is it logical to make ML like human brain?
- Instead of language based models would it be more likely to construct a model based on dominance hierarchy ?
- What mimicking spike trains actually do? How can they be useful? Would HTM applications have dangerous uncontrollable outcomes? numenta
- Brain is too complicated to build up bite sizes pieces spiking patterns, proteins and knowledge of corresponding genes play role in any disease processes. How to decide what to model and How much of intelligence is needed to understand the brain?
- Would modelling unhealthy behavior of brain work in ML? Full connectivity is never seen in the brain...
- Energy states of brain is also too complicated to model for example epileptic seizure state of brain is the lowest entropy state when its fully unconscious, full conscious state of the brain has the highest entropy.
- Level of modelling is also very important because, when we start to model brain at ion level would it be feasible? Deep and great understanding of neurons vaguely realistic neuronal understanding.
- Is modelling spike trains an efficient way of processing information? Are there any other methods? Would it be possible to understand and model consciousness based on spike activity modelling? (More of a philosophical discussion)
- What would be next when we digitize human brain completely? (Concerns about human brain project outcomes)

- Exact mathematical correspondence of brains is too complicated. Except the hippocampus and grid cell study by DeepMind (accidentally), they did a good job modelling this area because there were a wide range of literature regarding to grid cell activity and grid cell computation is relatively straightforward. This is not the case for the rest of the brain, and structurally distant areas work together in complex tasks so new mathematical definitions are necessary to make low dimensional processes to represent the higher ones.
- Quantitative connection of neural activities to model attention âĂŞ computational problem or not behavioral assessment problem or low level of abstraction âĂŞ super prototypical tasks to implement far from real brains since we only have some intuition.
- Taking neuroscience work does not necessarily mean useful or correct or applicable approximation of nature. Physics successive theory - theory is not always good for solution.
- Modelling automated tasks is different than conscious ones. There is no physiological or other limits-definitions between them.

Areas ML and Neuroscience are improving reciprocally: ML could help to achieve some understanding on levels of abstraction learning processes in the brain and it can also help to get true information from functional MRI representations in bold signals. It has applications with Calcium imaging for activity of human neurons could be tracked and psycho-physic information could be modelled.

Open Questions

- Since brain activity is mostly based on spike trains, what would be possible outcomes of closely mimicking this activity, ie Numenta?
- How to define and model consciousness/unconsciousness?
- Brains are not engineered to be most efficient their evolution provided some efficiency but it may not be most possible well established solution.
- It is claimed that human learn efficiently but how many parameters does human brain use during learning process?
- Could anyone neuromorphic computing could improve in ML?

Takeaways

- Take ideas from neuroscience and inspiration come from higher level applications rather than neurons.
- Fast learning systems etc buffer and slow learning system to change that abstraction.
- Neuroscience inspiredness âĂŞ attention conceptual different versions of attentions
- How to rank understanding... understanding is hard to quantify difficult to define
- Breakthrough comes from interaction of different disciplines and our discussion converged where philosophy and neuroscience intersects Hinton.

2.1.2 ML for Climate (and other Sciences)

Session Moderator: **Soukayna Mouatadid**, University of Toronto
Session Note-taker: **Tegan Maharaj**, Montreal Institute of Learning Algorithms

Slides: soukayna mouatadid

The session discussion included the need to understand the world better and to have more accurate predictions. Smaller resolutions are harder to achieve. But also we might not need resolutions for longer term (We already know the climate patterns well enough to understand climate change, just not to predict something at a 10km or week-long scale.).

Climate models are inherently average the chaotic system beyond a couple weeks. People care about things at a human scale (What is going to happen in my field next week?, year, 10 years?) How to couple large-scale models? It is necessary to ML model to spot mis-calibrations or anomalies. Get huge numbers of sensors.

Semi-Supervised prediction of extreme weather events

- Take output of IPCC climate model, like a big âĂŸmovieâĂŹ of the climate, predict labeled extreme weather events using deep 3D CNNs (e.g. hurricane etc.) Knutti and Sedláček [2013].
- IPCC is based on first-principles mathematical equations (so that things don not get into impossible states).
- Translate to ML: âĂIJconstrained to the manifold of physically possible phenomenaâĂİ
- How to do this in real time? (feed sensor measurements through climate model and then use deep prediction model maybe?).
- In general with physics models etc., use the physics model as âĂŸfeaturesâĂŹ for a DL model.
- Note that main purpose of climate models is to understand dynamics and behaviour and counterfactuals (main purpose is not to predict); canâĂŹt do this with a black-box predictor.

Cloud modeling

- Different resolutions; large-scale climate models and equations describing sub-grid processes of clouds; large-scale climate cannot resolve subgrid processes. Need data about climate before 1960 etc.; use a GAN to generate the data.
- General problem: putting together first principles understanding with empirical data
- Climate: increasing understanding through models, care about causal models / interventions Impacts and adaptations: care about predictions, not so much about causality
- What does interpretability mean for climate models?
- Being able to perform counterfactual interventions and see the result
- For interpretability attention maps are used but they would not be providing a global phenomena but they seem to be useful.
- Representation learning (how do you search for a representation under which the attention maps are focused (sparse/spatially compact) - can actually do this! We should do it!)
- Project on another model that you know about, e.g. physical models say how it relates to SDP electron rings.
- Schrodinger model: electron in real-space is everywhere, but spectral is in a specific place CNNs are not right for spectral data because of location in-variance? Temporal convolution makes more sense.
- ML and communities: Better to have something that the community understands than to have the most efficient thing.

2.1.3 Diversity and Inclusion

Session Moderator: **Cody Coleman**, Stanford University
Session Note-taker: **David Kale**, Netflix

Scope:

- Mix of folks with different concerns on teams about implementation of DandI:
  - ⋆ how it affects them personally
  - ⋆ how it affects their product or output

Topics:

- Negative side effects of diversity and branding
- Open discussion can normalize negative behaviors, e.g., saying that everyone has subconscious biases can make folks feel like it is OK to have them.
- Diversity branding initiatives can lead to empty symbolic gestures, e.g., donating to outside org but making no internal changes.
- How to wrestle with structural barriers?
  - ⋆ Financial problems
  - ⋆ Unfamiliarity, e.g., non-privileged folks at State Universities do not know that they should do research during undergrad!
- Problems at Stanford
  - ⋆ No community.
  - ⋆ When you are alone, you get a spotlight.
  - ⋆ When you are alone, toss responsibility on you.

Concern: ***how to wrestle with the notion that diverse candidates are necessarily less qualified?***

- Need to address pipeline at its source.
- Target hiring at non-elite schools.
- Remove superficial filters.
- Must deal with specifically intentional discrimination.

How do you get an organization to care about DandI?

- Find mentors!
  - ⋆ Not necessarily go to biggest name (mostly they are the busiest person).
- Need allies!
  - ⋆ They can speak up with fewer personal consequences.
  - ⋆ Sort of an âĂIJonly Nixon can go to ChinaâĂİ effect.
  - ⋆ Where are attendees from over-represented groups? – most attendees at meetings like this are from underrepresented groups.

Definitions:

- Diversity: invited to dance.
- Inclusion: asked to dance.
- Belonging: dancing like no one is watching.

Follow-up: diverse hiring is not enough

- Yale example: boosted diverse admissions to STEM but then disproportionate attrition

- At tech companies, disproportionate attrition in underrepresented groups

What good is diverse hiring if we send people into situations that traumatize them?

- Build communities and safe spaces for people to support each other.

Build organizational values into, e.g., job descriptions.

- Eliminate degree requirements that are inessential but may cause people to self-select
- Individual teams must take responsibility for addressing themselves.
  - ⋆ Think about supporting candidates once indoor.
- Description of overall system, specific numbers at each critical juncture.
- When you can measure and report something, you can (roughly) gauge whether it is getting better or worse in response to different interventions.

WhatâĂŹs a bigger problem: active malice or oblivious thoughtlessness!.

- Micro-aggressions
- Real problem is small, consistent, mindless behaviors that happen every day.

âĂIJBlack in AI is not a sideshowâĂİ.

- Need to elevate these issues to be a first class citizen.

How can DandI initiatives backfire?

- How can we lift the load from shoulders of underrepresented folks?
- Allies (and big companies) can assume the logistical work.
- Put willing underrepresented folks in prominent positions to ensure authenticity.
- WIML, Black in AI, LatinXAI puts women in positions of leadership but has open volunteer call
- Connect to sponsors.
- Help with paper reviews.
- DonâĂŹt wait to be invited to help – reach out!
- How to get allies involved in non-paternalistic way?
- âĂIJThe majority of people are good but most of them need reminding sometimes.âĂİ
- How can we build in reminders?
- Codes of conduct in places like WIML with calls to action for attendees, especially allies, to self-enforce.
- Some things are beyond our control.
- Visas: probably beyond control of organizers but can mitigate by hosting meetings in diverse locations.

References/Further Reading

- McKinsey and Lean In. Women in the Workplace Report mckinsey
- Clarke and Clegg. Changing Paradigms Clarke
- Myers. Moving Diversity Forward: How to Go From Well-Meaning to Well-Doing. Myers
- Myers. What if I Say the Wrong Thing?: 25 Habits for Culturally Effective People. Covey [2013]

2.2 Day 2 Session 2

Session 2 took part between 1-2 pm. The session topics were *Generative Models for RL*, *Causal Inference, especially for small-data regime*, *Strategies to Reduce the Computational Cost*, *Ethical/Responsible AI Development*, *Building and Monitoring Production-Ready ML /Distributed Training* and *ML in Medicine/Medical Imaging in TF*.

2.2.1 Causal Inference, especially for small-data regime

Session Moderator: None
Session Note-taker: **Jessy Lin**, Google Research & MIT Computational Cognitive Science Lab

What are causal models?

- Interventions over conditioning. Allows you to figure out causal effect from different factors, evaluate counterfactuals (âĂIJwhat ifâĂŹsâĂİ).
- Two main schools of thought: Rubin causal models Alaa and van der Schaar [2018] vs. Judea PearlâĂŹs do-calculus Pearl and Mackenzie [2018]

How to do causal inference, generally?

- Natural experiments: performing interventions / controlling the data collection process
- Conditioning-based methods: probabilistic methods for doing inference on graphs
- Sensitivity analysis: estimate the effect of assumptions we make about the model (causal inference makes strong assumptions about what cofounders are relevant)

How do we do causal inference when we only have observational data?

- Discovering Causal Signals in Images Chen et al. [2018], Nonlinear causal discovery with additive noise models Chen et al. [2013]
  - ⋆ Some signature in the noise that indicates direction of causality (âĂIJ Learn what the signature is from known causal structures. ")
- The problem with observational studies is that you can always show that two graphs with opposite causal directions produce the same observational data. We can only ever show that it is highly likely that the causal direction is a certain way / does not occur in âĂIJnaturalâĂİ causal structures.
- Reference to example from Judea PearlâĂŹs book on historically how we showed that smoking causes cancer, showing that the causal structure where there is a gene that causes both smoking and cancer (rather than smoking causing cancer) is just highly unlikely. A good example of a case where it would be hard to do randomized controlled trials.

Applications and domains where this is useful?

- Holy grail of generalization is zero-shot transfer to other domains with similar causal structure Kilbertus et al. [2018] and other work on transfer learning in the Bernhard ScholkopfâĂŹs group Huang et al. [2018]: while distributions may change, underlying causal mechanisms don not change.
- Fairness, explainability / interpretability
- Notion of causality is natural in time series, but what about images? What is the difference between causality and something like score attribution?
  - ⋆ Attribution as learning what is causal to your model, but not necessarily what is causal in the world. E.g. learning wolf vs. husky based on snow in the background. There are still real causal mechanisms in the world to be learned.
  - ⋆ E.g. learning what makes a tricycle a tricycle is 3 wheels, not that it is near a child, even though every picture of a tricycle in your data-set might be near a child.
  - ⋆ Not going to generalize well unless you learn the causal mechanisms
  - ⋆ There is something deep here...in time series cases there is a notion of physical causality, but this is may really what is happening with image labelingâĂę it is a human notion because we label these objects. âĂIJCausality as useful to usâĂİ.

- ⋆ Some physicists define time in relation to causality

Why do we care? What do people want to do with causal inference?

- Zero-shot learning and generalization data-sets?
  - ⋆ In healthcare at least, comes from domain knowledge
- Examples of learning the structure itself?
  - ⋆ Mostly learning weights, but structure learning is a growing subfield
- Axiomatic attribution for deep networks Mudrakarta et al. [2018]

Random questions and ideas

- How do causal graphs model echoes?
  - ⋆ Bayes nets do not include loops.
- Any examples of RL and causal inference? The equivalent of âĂIJexperimental dataâĂİ in ML.
  - ⋆ Taking actions is the equivalent of interventions.
  - ⋆ Extracting the implicit causal model learned by an RL agent? We do learn our actions have effects on the world.
  - ⋆ Intrinsic social motivation via causal influence in multi-agent RL: incentivize agents to have causal effects on other agents Grudin and Jacques [2019].
- Causal inference is strictly harder than what we usually need, because we are trying to estimate a whole class of distributions. Idea from Bayesian ML that we do not need to learn the whole posterior for prediction – could the same be true here?
- Examples of using causality to validate our models. If we know the causal structure, making sure what a model has learned is consistent with that knowledge.
  - ⋆ Some work on this in interpretability.
  - ⋆ Examples where humans are bad at causal inference.
  - ⋆ Luck and superstition (see studies on pigeon superstition when given stochastic rewards)
  - ⋆ Counterfactuals in theory of mind, diplomacy, social interactions.
  - ⋆ Way we form narratives post hoc, teleology of technology, anthropic principle.
- Hebbian learning rule as a âĂIJprior over causalityâĂİ.
- Examples of causes that seem to happen before their effects.
  - ⋆ âĂIJAcausalâĂİ e.g. stock prices; peopleâĂŹs buying behavior will depend on what they think will happen in the future.
  - ⋆ Cause must always precede effect, but observation of effect may precede observation of cause.

Resources

- Tutorial on causal inference Kiciman and Amit
- Judea PearlâĂŹs Book of Why: history, motivation, and intro to causal inference Pearl and Mackenzie [2018]
- Discovering causal signals in images Chen et al. [2018]
- Nonlinear causal discovery with additive noise models Chen et al. [2013]
- Learning independent causal mechanisms Kilbertus et al. [2018]
- Axiomatic attribution for deep networks Mudrakarta et al. [2018]
- Intrinsic social motivation via causal influence in multi-agent RL Grudin and Jacques [2019]

2.2.2 Building and Monitoring Production-Ready ML/Distributed Training

Session Moderator: **Cody Coleman**, Stanford University
Session Note-taker: **Miti Modi**, University of Toronto

Number of Attendees: 18

Brief information about the session

This session examined some of the common challenges of trying to build and deploy machine learning models into productions. Some best practices that are being used by practitioners were discussed.

Discussion:

Biggest challenges in building production-ready ML systems:

- Data is not clean and structured.
- Databases and systems do not talk to each other.
- Not every company has the compute power.
- How of them can do rapid prototyping in ML?
- You have to constraint people to a version sometimes, if there is an SLA to make a model available to peopleâĂŹs devices .
- Designing an integration test is really difficult âĂŞ How to assess whether a recommendation is good?
- When should the model be retrained ?
- 3rd party systems are pretty unreliable right now âĂŞ because the vendors are still trying to figure out what is needed.
- Is it necessary to invest on hardware?
- It can be necessary to design a model that gives some amount of control to the users - which may impact the model outputs.
- Reproducibility is bad. ML is a decade or 2 behind the software engineering in handling reproducibility.
- Tech stack for a Data Scientist vs for Production is very different : People who write the production ready ML do not know what the model is doing.
- Kubernetes without assistance is difficult to do for someone who has never done production ready ML (ie. for a data scientist) âĂŞ engineering teams are necessary.

Some solutions / Best practices / Future Directions Data

- Version control of data. This may be easier to do in certain fields vs others âĂŞ ex. Location of medical images are not going to get changed and or those images will not be augmented.
- Controlled experimentation to get feedback on what features may be useful.
- Sometimes a feature that is not as important may not be practical to use in production because it needs 3+ joins.
- Important to keep ETL (extract transform load) model in sync. Error messages currently do not explain what is broken in your pipeline and having this would be very beneficial âĂŞ eg. error message can specify that model was not trained on a new feature.

Building models

- Rapid prototyping for ML is difficult especially if you need A/B testing to evaluate the model.

Integration

- Can isolate dependencies with Docker âĂŞ so the library versions would not become a problem.
- Integration test for ML systems âĂŞ validation before it is used on people.
- Continuous integration helps âĂŞ infrastructure does not fully exist yet.

- ML community needs to define unit testing in ML and standardize by bringing best practices from software engineering to ML.

Model Retraining

- Detect input distribution drift (tools like Seldon).
- Velocity of change in production can help in identifying when to retrain or proactively retraining for better results and use stats to see if the model is actually better?
- Need a clear feedback loop then difficult to even identify that the model is not working .
- Automated retraining only when new data arrives or some teams also do weekly retraining without hyper parameters so that there are not large scale changes.
- Not a lot of changes to hyper parameter during retraining because what if it optimizes on something else starts making different decisions.
- People do not like too many changes in the product and hyper parameter changes during retraining may lead to such changes.
- When customers are enterprises/industries (doctors, bankers etc.)âĂŞ too many changes in model can be disruptive
- Less accurate model might be better than a model that keeps changing predictions.

Reproducibility

- Reproducibility on your production code âĂŞ should be able to run any version of the code on any dataset. This is important for accountability and retraining.

Data Scientists and Engineers

- Get great data scientists in management position and teach ML to great software engineers.
- Increase visibility of the production ready code to the data scientists so that they can interact with the code to ensure that the code is working correctly.

Cloud vs. Hardware

- Deciding the scale of the hardware is important and cloud is more flexible.
- Team may not be large enough to maintain hardware.
- If your experiments run for a long time then cloud could be more expensive.

Takeaways:

- There is a gap between data scientists who build the models vs. engineers who put them into production.
- Version control data for future reference is necessary
- Testing and integration of ML models need to be clarified and standardized.
- Lack of clarity amongst practitioners on when to retrain models âĂŞ stay proactive or wait for something to break? What if the model in production is helping medical professionals do their job? Fix it before waiting the breaking time of the model.
- More attention needs to be paid to reproducibility.

2.2.3 ML for Healthcare/Medical Imaging in TF

Session Moderator: **Marzyeh Ghassemi**, University of Toronto
Session Note-taker: **Neil Tenenholtz**, MGH & BWH Center for Clinical Data Science (picture notes), **Cynthia Habonimana**, California State University, GoogleAI (text from pictures),
Number of Attendees: About 50+

Brief Information about the session: This session discussed some of the common challenges and solutions regarding applying machine learning models in healthcare and medical imaging in TF. The session was divided into four sub-sessions due to a high number of participants considering the focus of expertise .NLP, Time Series Analysis, Medical Imaging and Implementation.

Subsession: Implementation *Problems of Implementation for ML in healthcare:* How to direct learning to get more information ?

Challenges

- Identifiability of data and its sources can be hard.
- Privacy: How to keep patient's privacy? How to get private data?
- Data sources can be diverse so how to handle data is one of the biggest problems.
- Are we observing the right things with the data, we have? Then, another question arises: Is the right data being collected?
- Interpretability (Important with recommendation) of data is the biggest challenge.
- How can humans understand?
- How do we make judgment? How can an expert judgment interfere the results?
- Consider Medical Facts: focus on extreme cases, also regression to mean.
- Softwares built for clinical workflow, this means implementation should be adaptable for this workflow.

Solutions

- Established justification procedures necessary.
- System has some flaws and any model(or human expert) can not be 100% correct all the time.
- Changing of the regulatory procedures GDPR for medicine but better ML applications.
- Make it easier for patients to get data - some patients want to share so encourage sharing.
- Hospital regulations must be adapted.
- Government should encourage AI implementation.

Subsession: NLP session

Challenges

- Data Access-permission issue, not all data is open-source and patients can be reluctant to share data.
- Problem Setting : Clinical vs non-clinical.
- Semi-structured language.
- Labels (e.g. CCS ICD 9, semi-supervised), label convention can be different.
- Defining outcomes is important otherwise it can be useless.
- Agreement on outcomes is necessary, what a clinician should expect on NLP for medical diagnosis? How a researcher can meet such expectations?
- Context from structured data must be inferred,
- Cross modal analyses on cross language settings is the biggest challenge.
- There are very few experts on practicing both medical research and NLP.
- External knowledge is necessary and tasks must be defined.
- Quality of core must be defined.

Subsession: Time Series Analysis

Challenges

- Sparsity, missing and biased sampling in datasets.
- Sporadic disease prediction is tricky.
- Unbalanced data, oversampled and undersampled representation.
- Patient heterogeneity and disease heterogeneity.
- There is a huge measurement drift.
- Access to datasets can be limited.
- Generability (new X).
- Meaningful inputs/targets may not be possible.
- Zero-Shot learning (new Y) .

Short-Term Solutions

- Sparse ending methods.
- Combining existing data.
- Matrix factorization.
- Class predict penalties may raise prediction ability.
- Subsampling can solve unbalanced data problem.
- Pseudo-alignment, generative modeling, transfer learning, standardized performance metrics, representation learning,manifold learning interpretability can also help to solve the challenges.
- Regulatory incentive.
- Causality needs to be well defined and can be used as diagnostic baseline.
- Anomaly detection require the human-in-the-loop.

Long-Term Solutions

- Passive data (normal, baseline) could make the problem a little bit easier.
- Patient self reports can be used to increase the amount of data.
- Passive and side information may help for prediction of sporadic diseases.
- Biomedical research can be used for confirmation along with clinical endophenotyping.
- Better devices necessary to handle measurement drifts.
- Engaging with patients may help to obtain more valid data.
- Change in the regulations is also necessary.

Subsession: Medical Imaging

Challenges

- Data access and obtaining data permissions is problematic.
- Data can be noisier than the acceptable threshold or may not provide the necessary information.
- Datasets are poorly organized, it can be hard to extract correct data for deep learning. Most datasets contain garbage rather than data.
- Modalities differ regionally and there is no common comparable standard.
- Datasets can be huge but does not contain enough data.
- Data is not publicly shared. It can be so hard to reach data.
- Quality / Quantity, generally data has low quality.
- Labeling is hard and requires expert knowledge.
- It is necessary to receive some consultation from clinicians to asess data.

- Differences in Scanners / Cameras / Imaging Data properties may result very differently. Additionally clinical processing pipeline differences result with different types of image perturbations.
- Resistance by doctors and institutions, not everyone agrees to share hard obtained data.
- Clinicians are generally lack programming experience and do not understand how AI can be helpful for them.
- Challenge with augmentation interpolation space, sometimes it is impossible to use augmentation, it is necessary to propose medical image specific methodologies.

Proposed Solutions

- Standardization of data storage formats and image formats are necessary.
- Methods of sharing data - Incentives.
- Explainability of model must be increased.
- Instad of augmentation generative models can help to obtain more data.
- Developing a quality filter could remove unwanted sort of data.
- More filters in first layers (CNN) so it can overcome noisy data.
- Domain specific architecture could be developed.
- One shot/Few shot learning methodologies can help to overcome the problem of lacking enough data.
- Incremental Training - proper output.
- Google Cloud Healthcare API provides a good standard.

References/Further Reading

A Survey on Deep Learning in Medical Image Analysis Litjens et al. [2017]

2.3 Day 2 Session 3

Session 3 was the last session of the SOCML and took part between 2:45 pm to 3:45 pm. The session topics were *AGI/Alternatives to the Reward Maximization Framework*, *Building successful AI Startups*, *Reproducibility*, *Role of ML in Non-tech Industries*, *Differential Privacy* and *Limitations of the GAN framework*.

2.3.1 AGI / Alternatives to the Reward Maximization Framework

Session Moderator: **Steven Stenberg Hansen**, Stanford University
Session Note-taker: **Ashley D. Edwards**,Georgia Institute of Technology

Number of Attendees: 22

Brief information about the session:

- AGI $\rightarrow$ AI good at general purpose tasks
- In general setting, unclear where reward comes from
- Considering the reward outcomes, how tasks should be defined?
- Notions of goals do not always translate to rewards.

Key Topics Presented:

- Exploration and intrinsic motivation
- Having agents train each other
- Partially Observable MDPs
- Model-based learning
- Multi-agent RL
- General agents

Controversial Key Points

- If agent can not self-preserve then not intelligent,
  - ⋆ Argument: Not sufficient
  - ⋆ Argument: Rocks self-preserve
  - ⋆ Argument: Intelligent beings do not always self-preserve (e.g. parents will sacrifice themselves for their children).
- Cannot be intelligent without other agents,
  - ⋆ Argument: Still intelligent but might not be seemed in that way to the outside observer.

Discussions: Exploration and intrinsic motivation

- Maximize surprise
- Intrinsic motivation is not always curiosity:
  - ⋆ Suppose agent has generally negative reward but gets hungry
  - ⋆ As agent gets hungry, it is more likely to explore to avoid even though it is getting some negative things.
- Exploration that is not maximizing entropy
- Thomson sampling: Maximize reward w.r.t. belief
  - ⋆ Know different actions have different utilities
  - ⋆ Sample from distribution
  - ⋆ Can not do Thompson sampling if you only have intrinsic reward because it assumes you have some external reward.
- Bayesian optimal exploration
- May need to constrain curiosity.

Having one agent train another

- Imitation learning
- Hierarchical RL
- Might be easier to learn differently than how humans learned.
- Might be easier to learn with human-in-the-loop.

Partially Observable MDPs (POMDPs)

- RL $\rightarrow$ MDP (Memoryless)
- POMDP has state that is insufficient / needs history.
- Planning done in belief state
- Given belief about the state, what the action results in the most information?
  - ⋆ Can not tell difference between uncertainty of model and world.
  - ⋆ Need to have some idea of stochasticity.
  - ⋆ Coin flip vs myself
  - ⋆ One decreases if you learn more (If not learning then move on)
  - ⋆ Static TV: Agent in curiosity stares at TV. If there is background hum we ignore quickly. Do we need some way to filter out this noise?

Model-based learning

- Model-building based on how well you can predict curiosity.
  - ⋆ Causal curiosity.
  - ⋆ Causality can map into POMDPs (actions are different causal mechanisms).
  - ⋆ Twitter: Judea Pearl thread.
- Just learn giant model instead of reward
  - ⋆ Reward has to be specified over and over .
  - ⋆ Addresses need to get rid of reward.
- Drawn to tasks that might simplify world model.
  - ⋆ Allows connections to be made between different things in model
  - ⋆ Compresses world
- How much model-based RL will be necessary?
  - ⋆ Does it have to be model vs large value function?
  - ⋆ What is model and what does it include?
- Need model of humans.

Multi-agent RL

- Seems necessary for AGI.
- Is intelligence relevant without other agents? What sorts of interesting things can you learn?
- Most agents do not model other agents .
- Modeling preferences or goals would be output of societal equilibrium .

2.3.2 Reproducibility

Session Moderator: **Gideon Dresdner**, ETH ZÃijrich
Session Note-taker: **Khimya Khetarpal**, Montreal Institute of Learning Algorithms

There were about 15 attendees for this session. With more and more accelerated advancements in ML and in particular DL and Deep RL, reproducibility in research is a burning issue.

Key topics discussed:

- Why do I care about reproducible research ?
- What would the incentives be?
- Challenges in making research reproducible
- Ongoing efforts in this direction.

What are the incentives - Why would one care?

- Minimal restrictions – randomized selection of 10% papers – to be inspected / run / etc.
- Industrial/Academic tracks – academic tracks need to release the code. Academic vs industry track: which is very challenging as much of the accelerated research today is due to collaborations across academia and industry. Is there a solution which will not hinder science and still achieve fairness in reviews in terms of open source code being marked a plus?
- Some NLP conferences –> requires that you *submit an abstract.* much earlier than the deadline and you cannot change the claims during submission. Key idea being baby steps towards the vision.
- Data driven approach – once we are collecting code, we can start data analysis and see what the state actual is.
- Natural selection for tooling.
- Leverage students learning ML to reproduce experiments – reproduce conferences. (cite ICRL repro challenge). Broken incentive structure. Sort of treats newcomers as second class citizen to test and reproduce other people's code.
- Need some kind of point based system to encourage and impact the review process to encourage this.

Current challenges in reproducible research

- Randomization in experiments? Can we build that in? Reporting confidence intervals? If it is not robust to "some noise" then it is garbage anyway!
- Are cherry picked examples valid as results? Some authors have shown top 3 runs.
- Problems in evaluating RL algorithms. Inconsistency in doing so.
- Human evaluation vs automated evaluation. Which is better?
- Rethinking how we are evaluating in RL.

Efforts towards reproducible research:

- Data *version control*
- Hyper params version control
- Pachyderm? Netflix – storage is cheap so persist everything – data work flow startup. Open source.
- Single point of entry?
- Software package to track all the things. UUID.
- Lite Tracer.
- Repo to Docker
- Orion.
- Reevaluate.

2.3.3 Limitations of the GAN framework

Session Moderator: **Alexia Jolicoeur-Martineau**, Montreal Institute of Learning Algorithms
Session Note-taker: **Meltem Atay**, Middle East Technical University
Number of Attendees: 30

Brief information about the session

GAN models collapse, they tend to converge on nonsensical clear image producing in latent space. Generator gets the data distribution generator disguises itself into key distribution, back propagation into generators' value function.

Mathematically this is binary cross entropy loss mini-max minimizes loss to causes discriminator to mis-classify the sessions. Learned divergence: divergence depends on the value function learning the divergence by maximizing to train generative models. These could be main reasons of why training generative models are too expensive. Usage of maximum likelihood could be improved because taking distribution and minimizing is not enough to optimize training.

Since the session aimed to discuss limitations of GAN framework all the key points were mainly controversial at least no one claimed GAN framework is perfect as it is!.

Main Discussion Keywords

- Discriminator training
- Generator training
- Model evaluation
- Nash equilibrium
- Architecture
- Global inconsistency

How can the principle model select the good output? Evaluation of the methods such as inception score (FID) that measures the entropy. A realistic classifier flavor of argument based on inception score but this is a slow evaluation technique.

Optimization is the problem of the definition even the classifier is not able to reach the local minima. In the case of that gradient descent does not converge to Nash equilibrium it can be a good cross entropy measure. Nash equilibrium is not vitally necessary for producing the ideal output. Frechet Inception Distance (FID) is a good metric but it is not enough to define how GANs work correctly so we need more well-defined metrics.

Attention could be added to GANs to improve the generator results. GAN is based on such evaluation metrics so it can never replace a predictor. GANs generates data from the distribution of the input by imitation of the replicating same distribution.

The conditional probability of (Y|X) E(Y|X) is hard to estimate. GANs optimize the divergence based on examples and features of the different kinds of data. GANs work on a probabilistic way without using its well defined rules. They are unsupervised but you can estimate domain adaptation which is an interesting property of GANs.

Considering these key points, discussion went under two major questions:

How to Make Discriminator Better?

- Regularization of discriminator by penalty or other form of normalization is the emerging best practices of the regularization. Gradient penalty is doable but it is expensive. To make the gradient nonzero, spectral normalization of GANs penalty requires the scaling of the weight layer directly.
- Spectral normalization and Constraint optimization required.
- Principle components and full of variability of expressive network is necessary.
- Early stages of training is similar to approaches of lipschitz optimization as a form of convex optimization which eventually converge. This property explains how GANs add smoothing to the input objects and get inputs look like they converged on similar outputs.
- Adversarial robustness is a must Tsipras et al. [2018].
- Repeated vector multiplications begin to point towards to imposing loss direction of overwhelming transformations, other directions blow out multiple inputs and force to convergence.

How to Make Generator Better?

- Prerequisite of test data, different objective reverse KL diversion can result extremely bad. Evaluating divergence term is enough to show the combination of one works for one discriminator to another. Although GANs do not memorize FID has training set memorization, hence it is necessary to develop evaluation functions used in the test sets .
- New GAN evaluation metric= duality gap distance on equilibrium: 0 only if there is Nash equilibrium. Grnarova et al. [2018]
- Could GANs be used to evaluate GANs!?
- What makes training expensive? Initialization is important, there are speedup techniques, optimal generators will collapse. For generator one can apply closed form classifier of the nearest neighbor, how to do that? By saddle point optimization to lower the training loss.
- Unconstrained optimization problem: Training the discriminator to learn the divergence, evaluation metric should include perceptional correlation. In practice generator does not minimize that quantity, it is an optimization problem (unconstrained optimization) divergence is not the ideal one if you got so far on optimizing the discriminator. So it will see both real and fake data. This situation satisfies the definition of divergence, as soon as you change G and keep optimizing, changes of generator result with divergence changes.
- Simultaneous gradient descent is also not that simultaneous, so it can be logical to use separate loss functions from the beginning, hence simultaneous optimization is not achievable.
- Random noise helps but there are issues to fix Lučić et al. [2017].

Takeaways and Open Questions How to define memorization? It can be achieved by assigning to much probability to training set, so distribution from training set can be covered. Likelihood is not terrible but it has flaws on it. Such flaws imply that, it is necessary to have more than one metric evaluation of generative models.

Is there any applied GANS imagine the first GAN video? (1 day later it happened almost? (Almost because not a pure applied GAN video!)Nvidia

More on Duality Gap Metric: If you have a generator and a discriminator working together, they must be evaluated. The most possible adversarial generator has the largest value and equal to 0. In practice such measure can be used for non-biased estimate, to train GANs, for computation adversarial generation and to evaluate value function. Duality gap metric works well, natural curve of GANs, generator loss and discriminator loss oscillate and it is hard to tell when they stop. When model collapses it measures if it is close or not to Nash equilibrium.The metric is domain independent, correlates with embody simulation. When GANs not guarantee Nash equilibrium which may exist or not, it is assumed to be exist. Duality gap shows that how GANs correlates with disturbance level specifically when model collapses Grnarova et al. [2018].

3 Acknowledgements

We sincerely thank to Google AI team, Google Toronto and SOCML organizing committee and all the participants for their incredible contributions. Compilers team owes deep gratitude to all moderators and especially note-takers as well as the participants of SOCML. Since the compilation duty is not more than simple editing to represent the discussed ideas more clearly, all the credit of this article belongs to SOCML note-taking volunteers equally.

4 Additional Information

References are external reading of the cited corresponding discussion topics.

5 Notes for the future compilers

Instant messaging among note-takers and compilers while gathering notes and during editing would be useful. It would be better if note-takers do some editing on notes before delivering them to compilers. Compiler team could be divided into typing error inspector, editors (grammar and content control), text formatter and citation finders. Due to busy schedules of the schools or conferences afterwards of SOCML, it could be nice to share basic tasks with more volunteers (preferably more than 5 people with duplicates of each tasks). It is also necessary to keep the track and such management issues may produce delays on writing process of the report. Using keywords may be useful while taking notes but in many cases those keywords could be meaningful only for the writer so notes should be edited on readable format before compilation.

References

Khan Academy. Price elasticity of demand and price elasticity of supply. https://www.khanacademy.org/economics-finance-domain/ap-microeconomics/ap-supply-demand-equilibrium/ap-price-elasticity-tutorial/a/price-elasticity-of-demand-and-price-elasticity-of-supply-cnx.

Alessandro Achille, Glen Mbeng, and Stefano Soatto. The dynamics of differential learning i: Information-dynamics and task reachability. *arXiv preprint arXiv:1810.02440*, 2018.

Ahmed M Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

Daniel E Berlyne. Conflict, arousal, and curiosity. *American Psychological Association*, 1960.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training. *arXiv preprint*, 2018.

Zhitang Chen, Kun Zhang, and Laiwan Chan. Nonlinear causal discovery for high dimensional data: A kernelized trace method. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1003–1008. IEEE, 2013.

K Alec Chrystal, Paul D Mizen, and PD Mizen. Goodhart's law: its origins, meaning and implications for monetary policy. *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart*, 1:221–243, 2003.

Thomas Clarke. Changing paradigms: The transformation of management knowledge for the 21st century. https://www.amazon.com/Changing-Paradigms-Transformation-Management-Knowledge/dp/0002570157/.

Stephen R Covey. *The 8th habit: From effectiveness to greatness*. Simon and Schuster, 2013.

Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7111–7118. IEEE, 2018.

Sara M Freeman, Kiyoshi Inoue, Aaron L Smith, Mark M Goodman, and Larry J Young. The neuroanatomical distribution of oxytocin receptor binding and mrna in the male rhesus macaque (macaca mulatta). *Psychoneuroendocrinology*, 45:128–141, 2014.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100 (1-3):70–87, 2006.

Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Nathanael Perraudin, Thomas Hofmann, and Andreas Krause. Evaluating gans via duality. *arXiv preprint arXiv:1811.05512*, 2018.

Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307, 2012.

Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. *arXiv preprint*, 2019.

David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018. URL http://arxiv.org/abs/1803.10122.

hackadamia. https://www.hackademia.io/.

Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. Amc: Automl for model compression and acceleration on mobile devices. In *European Conference on Computer Vision*, pages 815–832. Springer, 2018.

Geoffrey Hinton. Machine learning for neuroscience. https://neuralsystemsandcircuits.biomedcentral.com/articles/10.1186/2042-1001-1-12/.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560. ACM, 2018.

McKinsey Global Institute. Visualizing the uses and potential impact of ai and other analytics. https://www.mckinsey.com/featured-insights/artificial-intelligence/visualizing-the-uses-and-potential-impact-of-ai-and-other-analytics/, 2018.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, pages 5936–5946, 2017.

Emre Kiciman and Sharma Amit. Causal inference. https://causalinference.gitlab.io/icwsm-tutorial/.

Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

Reto Knutti and Jan Sedláček. Robustness and uncertainties in the new cmip5 climate model projections. *Nature Climate Change*, 3(4):369, 2013.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

Victoria Krakovna. Gaming examples in ai. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

Yuxi Li. Rl in real life. https://www.dropbox.com/s/sedh13fh96gn4sy/RL_RealLife.pdf?dl=0/. Published 2018-11-30.

Yuxi Li. Deep reinforcement learning. *arXiv preprint arXiv:1810.06339v1*, 2018.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Mario Lučić, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

mckinsey. Woman in ml. https://womenintheworkplace.com/.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? *arXiv preprint arXiv:1805.05492*, 2018.

Verna Myers. Moving diversity forward: How to go from well-meaning to well-doing. https://www.amazon.com/Moving-Diversity-Forward-Well-Meaning-Well-Doing/dp/1614380066/.

numenta. https://numenta.com/machine-intelligence-technology/?gclid=EAIaIQobChMI74rl3Lfj4AIVE6WaCh0iyAkYEAAYASAAEgIbX_D_BwE/.

Nvidia. https://news.developer.nvidia.com/nvidia-invents-ai-interactive-graphics/?ncid=so-you-ndrhrhn1-66582E/.

Council of Supply Chain Management Professionals. What is supply chain management? https://cscmp.org/CSCMP/Join/About_Us/CSCMP/Join/About_Us.aspx?hkey=e15eb27f-d327-4ef3-89f9-2ade73e34a55/.

Chris Olah. Feature visualization. https://distill.pub/2017/feature-visualization/.

P-Y Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. In *Progress in brain research*, volume 229, pages 257–284. Elsevier, 2016.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704. ACM, 2006.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

Scott D Sagan. The limits of safety. *Princeton, NJ: Princeton University*, 1993.

Jürgen Schmidhuber. Adaptive confidence and adaptive curiosity. In *Institut fur Informatik, Technische Universitat Munchen, Arcisstr. 21, 800 Munchen 2*. Citeseer, 1991a.

Jürgen Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pages 1458–1463. IEEE, 1991b.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991c.

socml.org. Socml tentative schedule. `https://docs.google.com/spreadsheets/d/1n7nQ4x6Pku-2oB9CiEmO7Hxv7lSDTG-eBwv1CJN6OaQ/edit#gid=0/`. Accessed 2018-12-19.

soukayna mouatadid. Ml for climate: challenges and opportunities. `https://drive.google.com/file/d/1Q-SwsrqpvyTrf-iUcsBwai2B2RORzaMk/view/`.

Social Good Summit. `https://www.un.org/sustainabledevelopment/events/social-good-summit/`.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

William Whitney. Disentangled representations in neural models. *arXiv preprint arXiv:1602.02383*, 2016.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.