Deep Learning
Summary and Challenges

COMP4660/8420

Bio-inspired Computing: Applications and Interfaces

Overview

- Quick summary

- Challenges
  - Biomimicry
  - Social Issues
  - Practical Issues

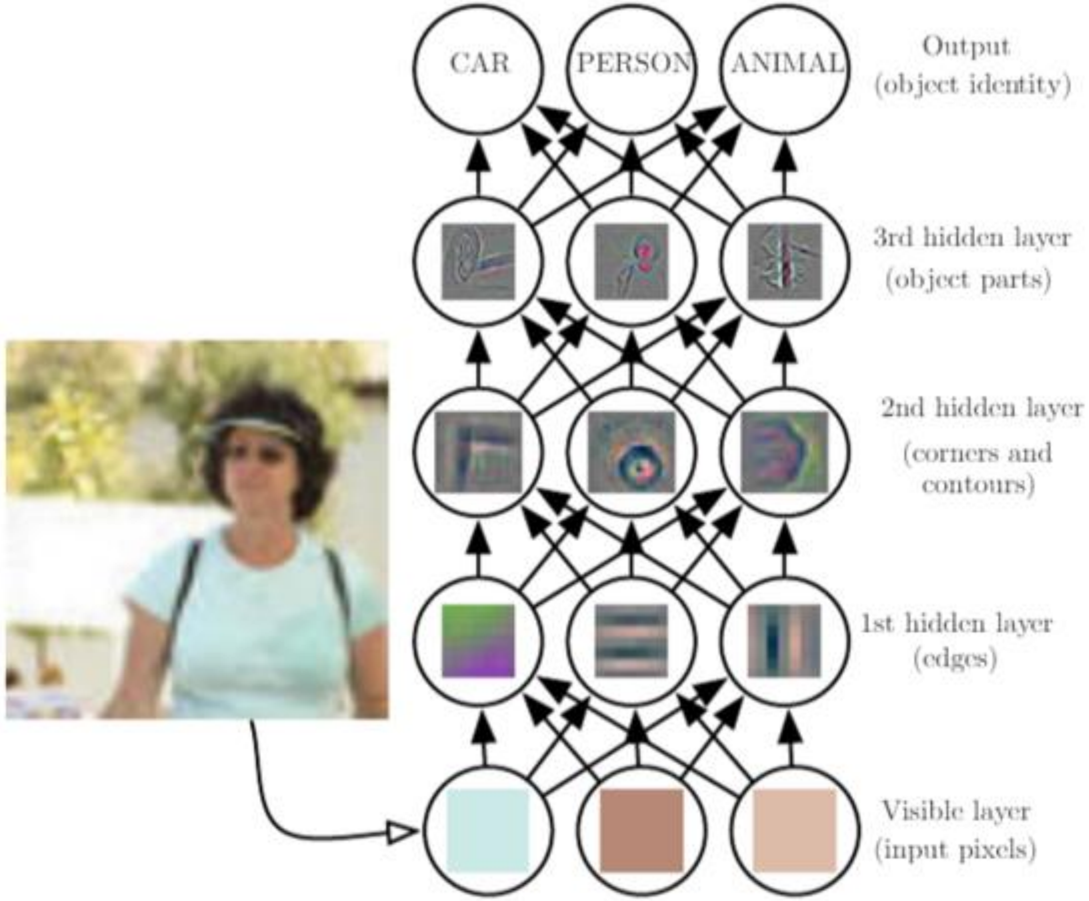
- Current and future directions

Deep Learning

A class of machine learning techniques that are able to understand the world through a hierarchy of concepts. Data representation is learnt through multiple levels of abstraction.

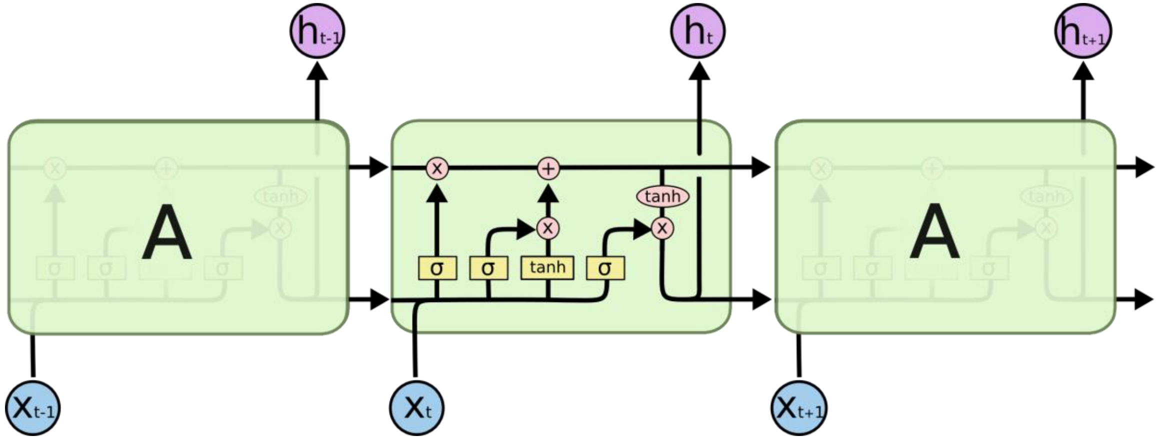
- Many biological inspirations
- Led to a shift from feature engineering to architecture design and hyperparameter tuning
- Made possible by availability of data, and advances in computational power (e.g. GPUs)
- Argued to suffer from a large amount of hype as a buzzword

Hierarchical Feature Extraction



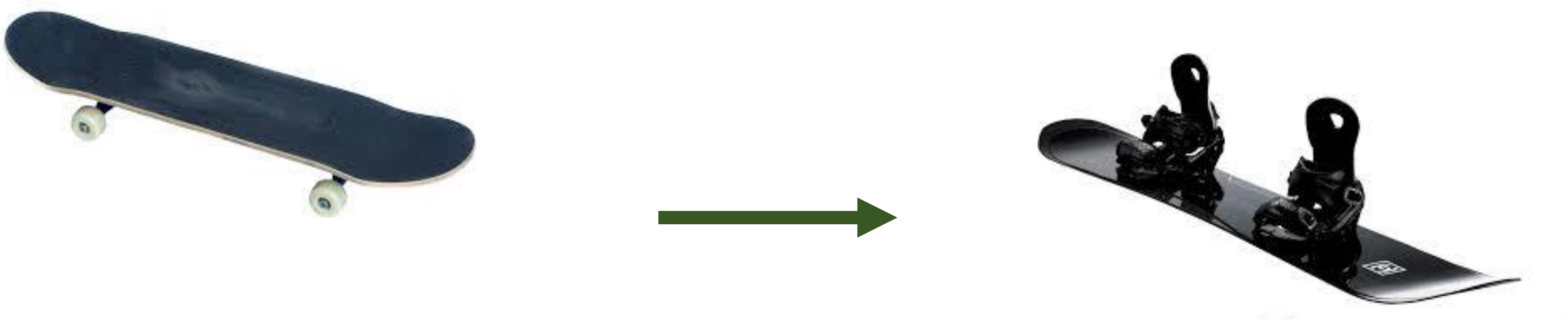
Sequential Memory

- Humans understand the world though an understanding of sequences and context
- Breaking down sequences into smaller units and remembering what happened previously to establish context



Transferable Knowledge

- Our brain is capable of using knowledge learnt previously and in other applications to help us solve new challenges



+ extra training

Biomimicry Introduces Challenges

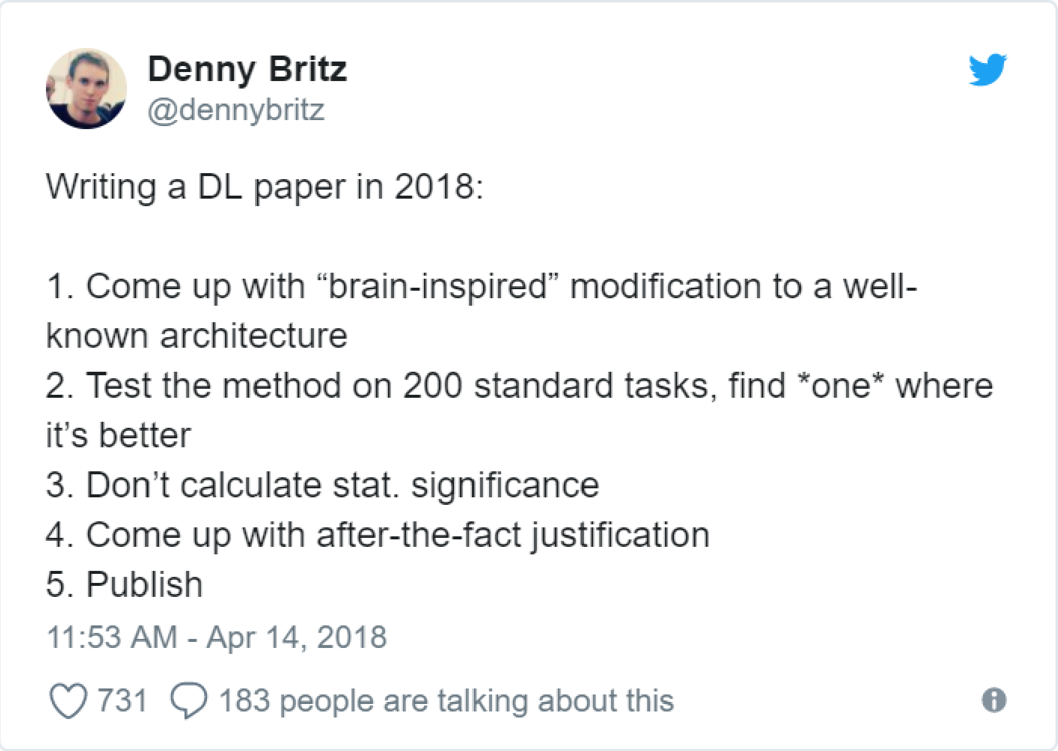
- In some of these cases, deep learning models (and other areas of machine learning) do a good job of applying biological understanding to mimic nature

- However, this is not always necessarily a good thing, as nature doesn’t always get it right either

- Furthermore, fundamental differences remain which have meant performances truly on-par with the human brain remain elusive

Toward Artificial General Intelligence

- Despite good progress being made from the use of biologically-inspired techniques, deep learning is still not a (complete) solution to solving artificial general intelligence
- Research is often caught up in identifying marginal performance gains on a single specific application without adequate scientific rigor (see Winner's Curse? On Pace, Progress, and Empirical Rigor)



Denny Britz
@dennybritz

Writing a DL paper in 2018:

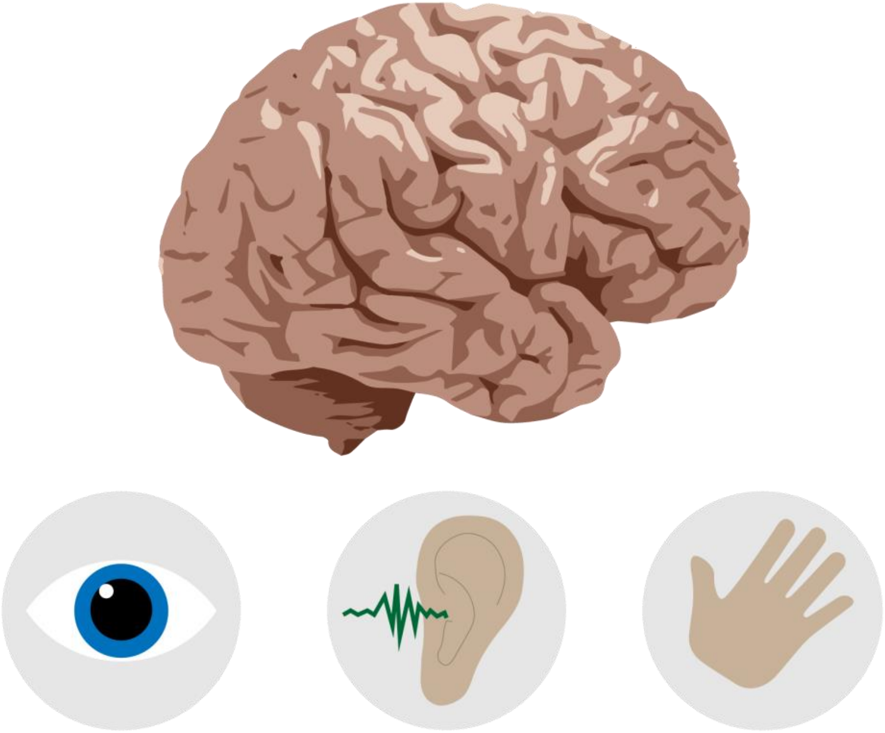
1. Come up with "brain-inspired" modification to a well-known architecture
2. Test the method on 200 standard tasks, find *one* where it's better
3. Don't calculate stat. significance
4. Come up with after-the-fact justification
5. Publish

11:53 AM - Apr 14, 2018

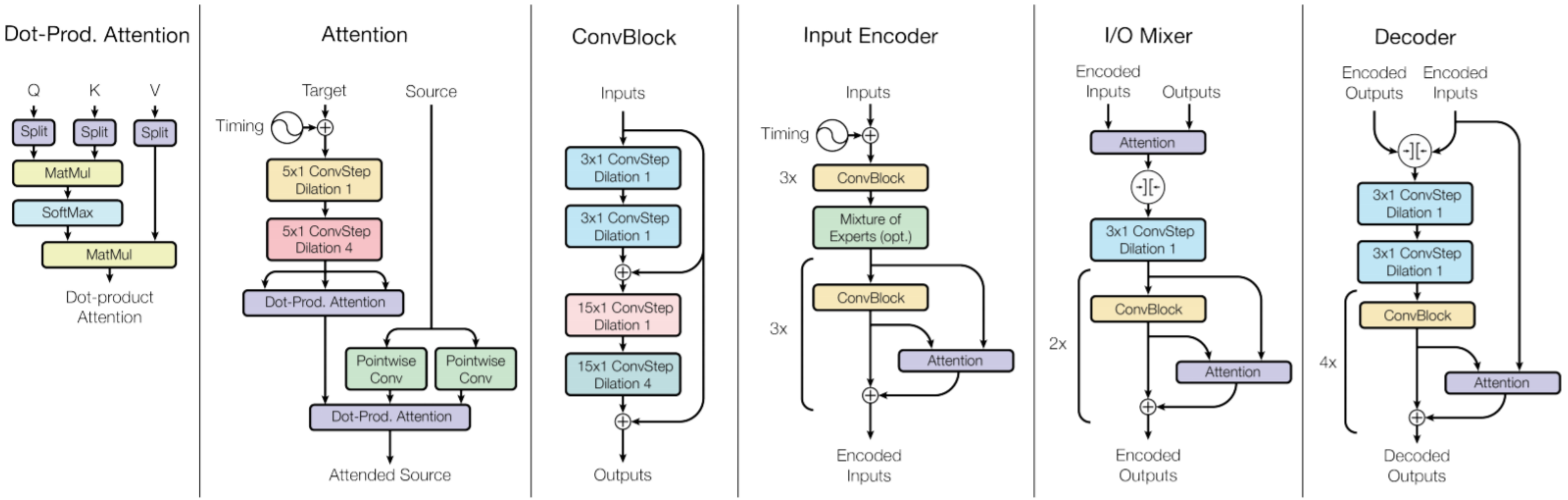
♡ 731 183 people are talking about this

Single network

- The brain performs multiple complex functions through a single network structure (albeit different neural pathways)
- Techniques have mostly focused on regularisation to generalise networks (eg dropout), which improve situational performance, but may reduce the capacity for dedicated cross-domain applications
- Researchers have been interested in seeking a single, unified network that is capable of performing many tasks (see *‘The master algorithm: How the quest for the ultimate learning machine will remake our world*’)

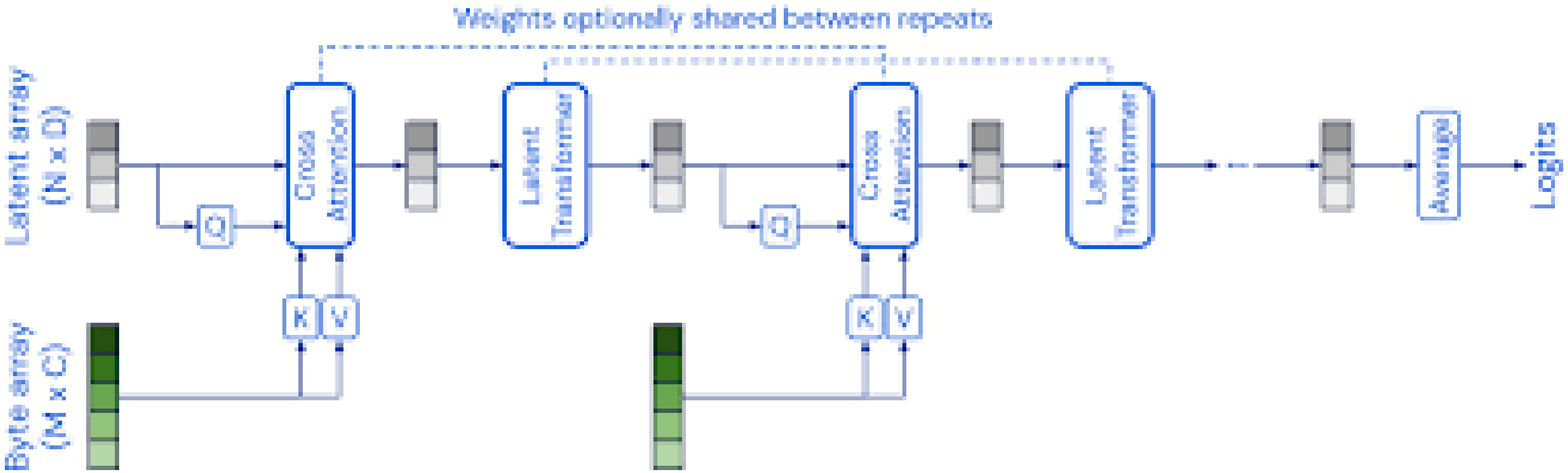


MultiModel



- From ‘One Model To Learn Them All’
- Computational blocks from multiple domains are incorporated into a single network, and trained on several different tasks.
- Schmidhuber theorised similarly in his recent paper ‘One Big Net For Everything’

The Perceiver



- From ’Perceiver: General Perception with Iterative Attention’
- Cross attention is used to incorporate information from any input modality into the latent vectors.
- Self attention mixes representations between latent vectors.

Social Issues

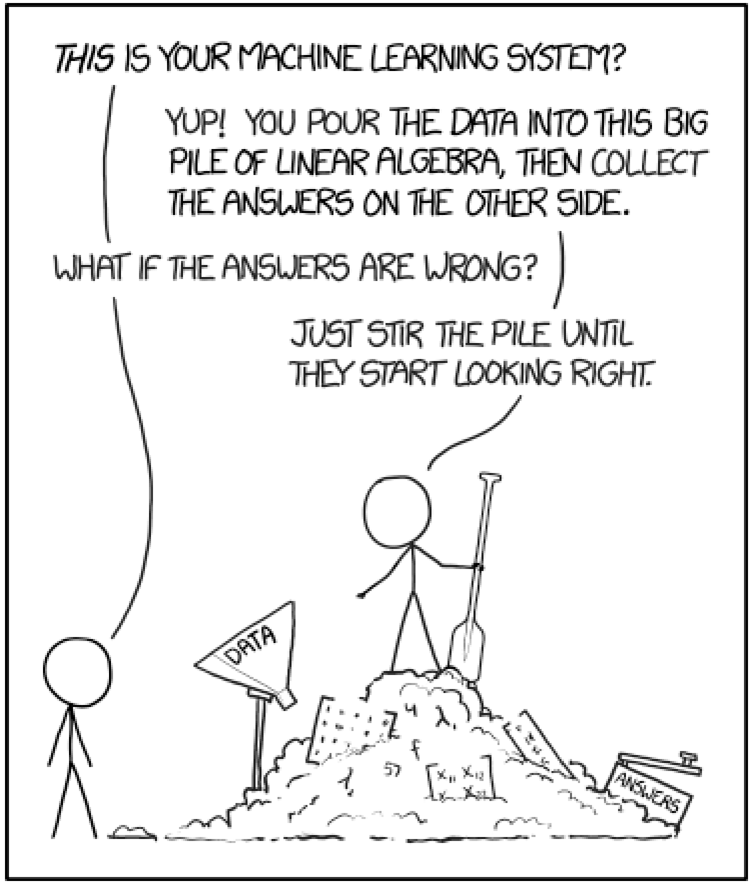
Social Issues

- Deep learning has exacerbated a range of social issues within AI research

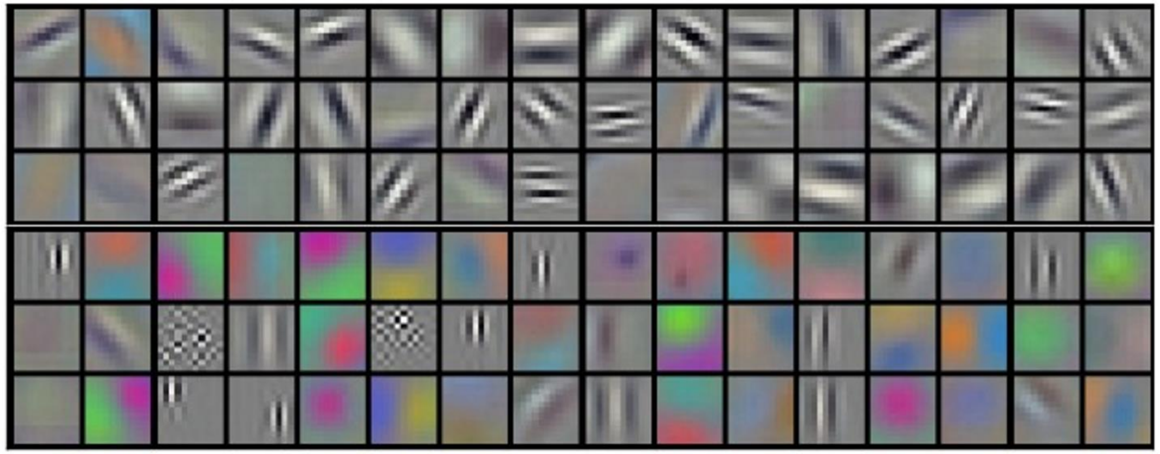
- Trust – how can we be sure that the output is accurate?
- Security – how can we be sure that someone can't maliciously poison the network?
- Ethics – how can we be sure that the network is operating in an ethical manner? Fairness

Black Box

- It is still not always entirely clear what makes certain deep learning models successful
- Models have millions of parameters so it is difficult for us to pinpoint exactly what conditions cause ideal outcomes
- There is a lack of feature introspection


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Feature Introspection

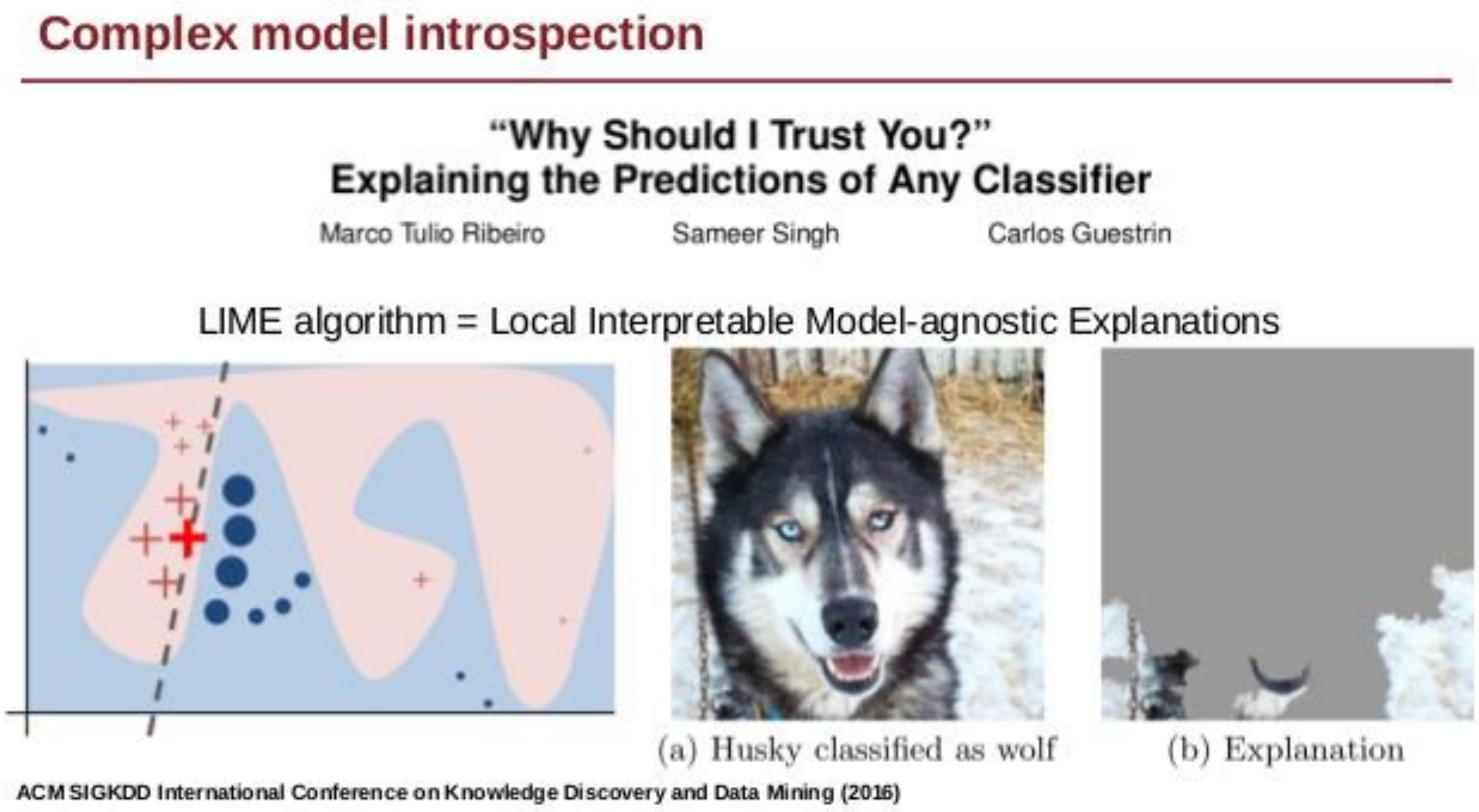


- As we don’t specify what features are learnt (only how many), we often do not have insight into the role that data features have in helping the network make a decision
- Many visualisation techniques allow researchers to try and understand what feature representations are being learnt for each layer (e.g. see ‘Visualizing Deep Neural Network Decisions: Prediction Difference Analysis’)

Interpretability

- Models need to be interpretable by humans in order to be useful
- https://christophm.github.io/interpretable-ml-book/
- Lipton (2016) proposes several factors for interpretability:
  - Simulatability – a human should be able to take the input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction
  - Decomposability – each part of the model has an intuitive explanation of purpose and how it works
  - Algorithmic transparency – understanding of the 'shape' of the error surface, and the impacts of training on convergence
  - Post-hoc interpretations through visualisation or explanations (including with examples)

Interpretability



Complex model introspection

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro Sameer Singh Carlos Guestrin

LIME algorithm = Local Interpretable Model-agnostic Explanations

(a) Husky classified as wolf

(b) Explanation

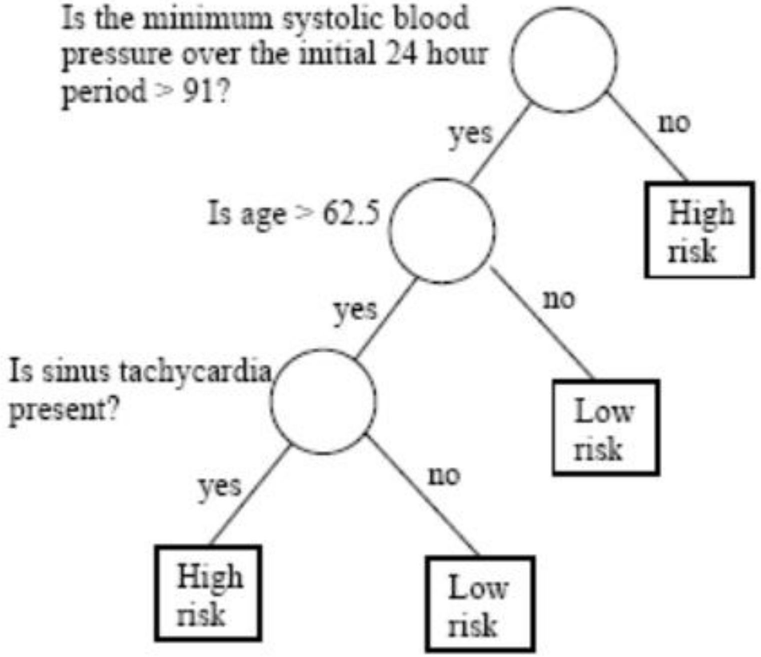
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)

https://arxiv.org/pdf/1602.04938.pdf?mod=article_inline

Trust

- A black box is not explainable to customers or the CEO; there is often a need for liability and accountability

- For example, a doctor making a significant medical diagnosis needs to be confident that a deep learning model produces accurate results. The doctor needs to be able to trace each step to understand and verify how the model produced its prediction


Is the minimum systolic blood pressure over the initial 24 hour period > 91?
yes
no
High risk
Is age > 62.5
yes
no
Low risk
Is sinus tachycardia present?
yes
no
High risk
Low risk

Security

- An aspect of trust is ensuring that the results are accurate and based upon well-formed data examples

- Theoretical and practical weaknesses in deep learning models mean they are vulnerable to attack, which is a problem as society becomes increasingly reliant on them

- Many deployed machine learning applications allow public access as part of its application

- Organisations need to be confident that deep learning applications remain true to their intended purpose and performance

Deep Learning Attacks

- There are many different attacks possible on machine learning systems, including:
  - Extracting private or sensitive information (from both users and the network itself)
  - Fooling classifiers with modified inputs
  - Corrupting the model with malformed inputs

- Many of these attacks exploit the on-line nature of deep learning systems that are now embedded in commercially-available and public systems

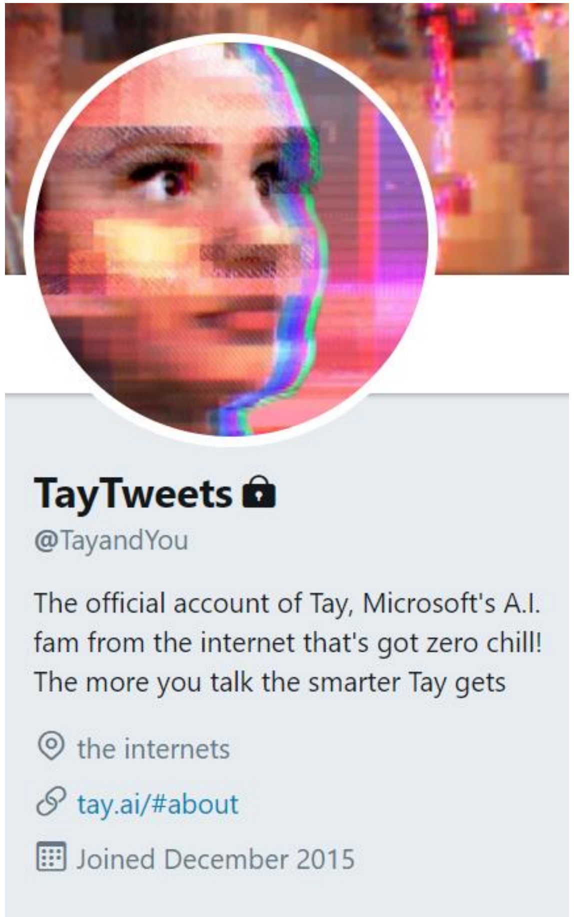
- More details regarding these can be found in ‘SoK: Towards the Science of Security and Privacy in Machine Learning’

Poisoning

- Manipulate the data used for training to confuse the model enough to prevent convergence, or to mislabel data to produce inaccurate outputs

- Many on-line systems constantly learn from new data given to it by real-world users

Example – Microsoft Tay



TayTweets

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets

tay.ai/#about

Joined December 2015

"As many of you know by now, on Wednesday we launched a chatbot called Tay. We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay. Tay is now offline and we'll look to bring Tay back only when we are confident we can better anticipate malicious intent that conflicts with our principles and values."

Peter Lee
Corporate Vice President, Microsoft AI + Research

An example of Tay's less-than-desirable tweets can be found in this article

Adversarial Attacks

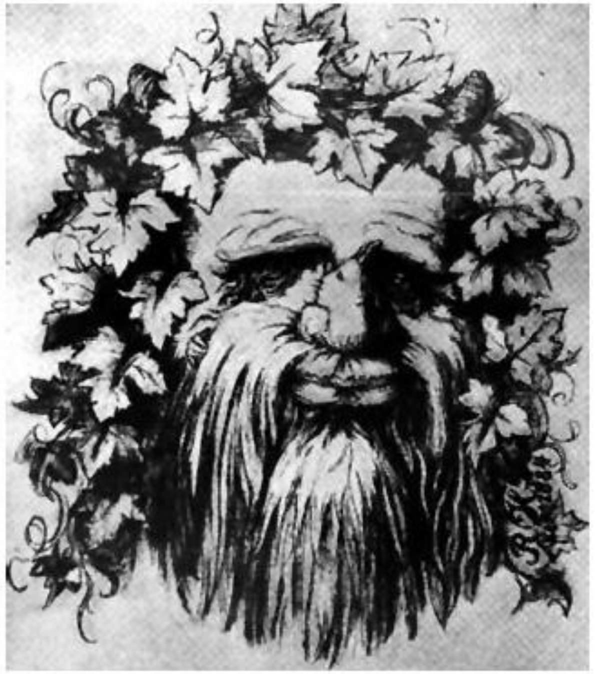
- Coerce a neural network to produce an output that is unintended.

- An attacker perturbs (modifies) an input that causes the activations of nodes in a neural network to cross classification boundaries or traverse steep regression gradients (see ‘Practical Black-Box Attacks against Machine Learning’)

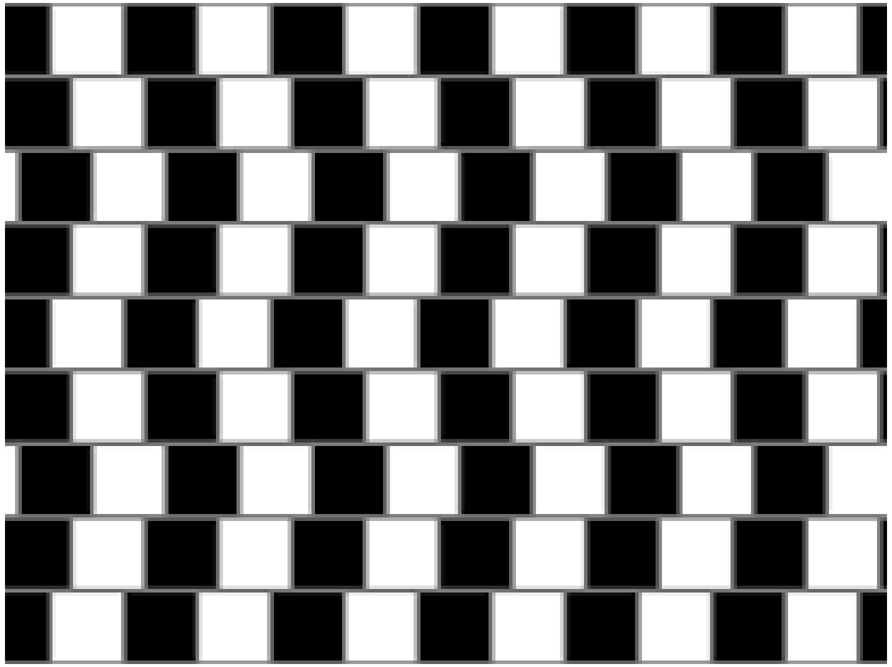
- Perhaps a biological limitation, akin to an optical illusion

Fooling The Brain

Old Man or two lovers kissing?

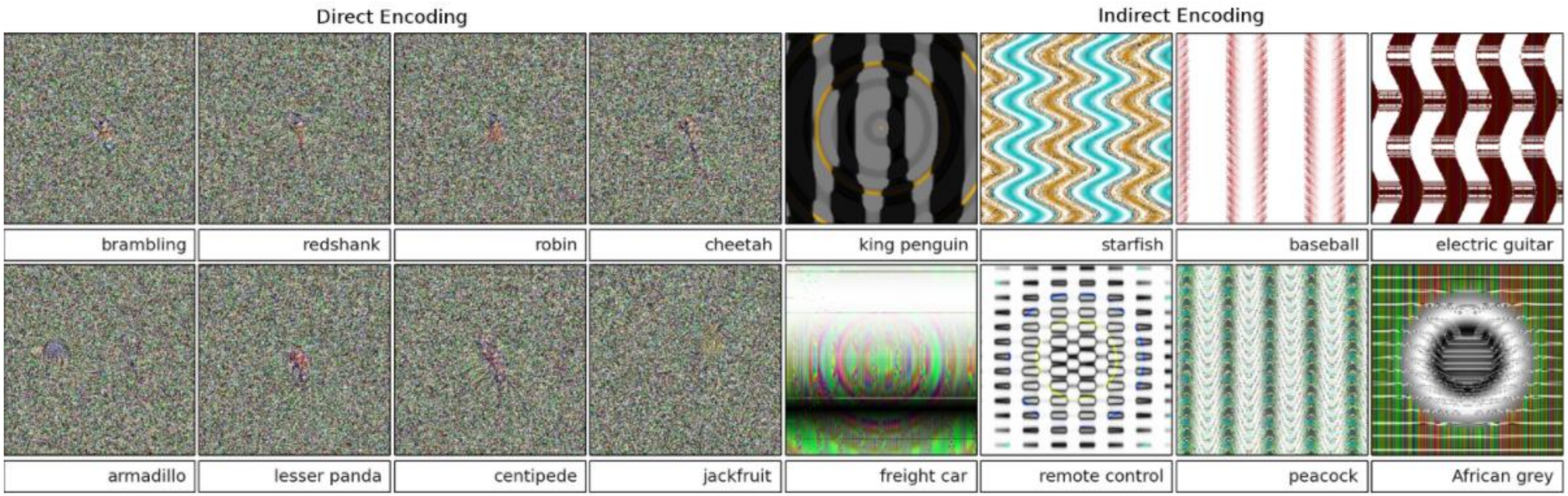


Parallel lines



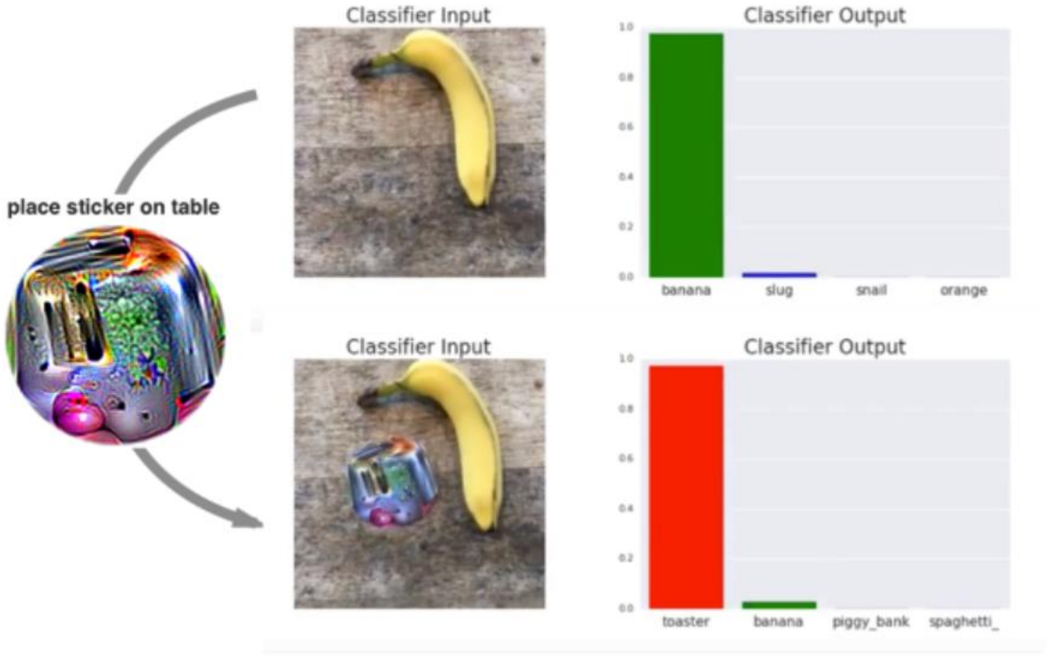
Deep Learning Interpretations

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

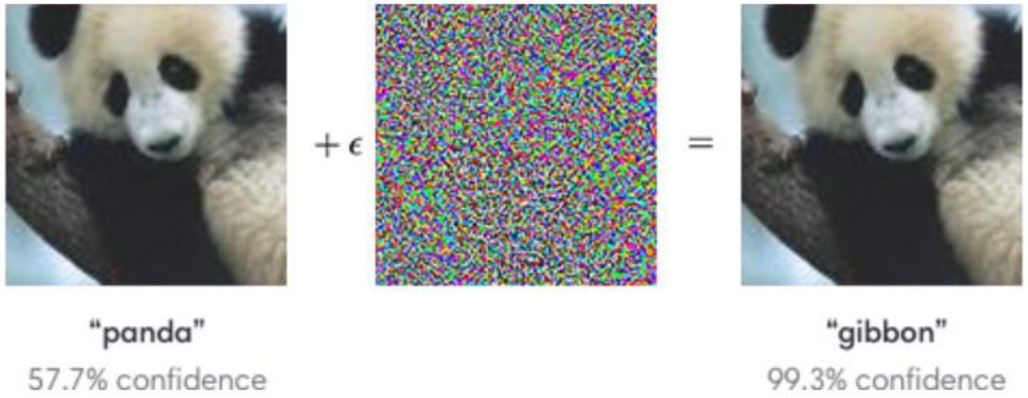


Fooling a Neural Network

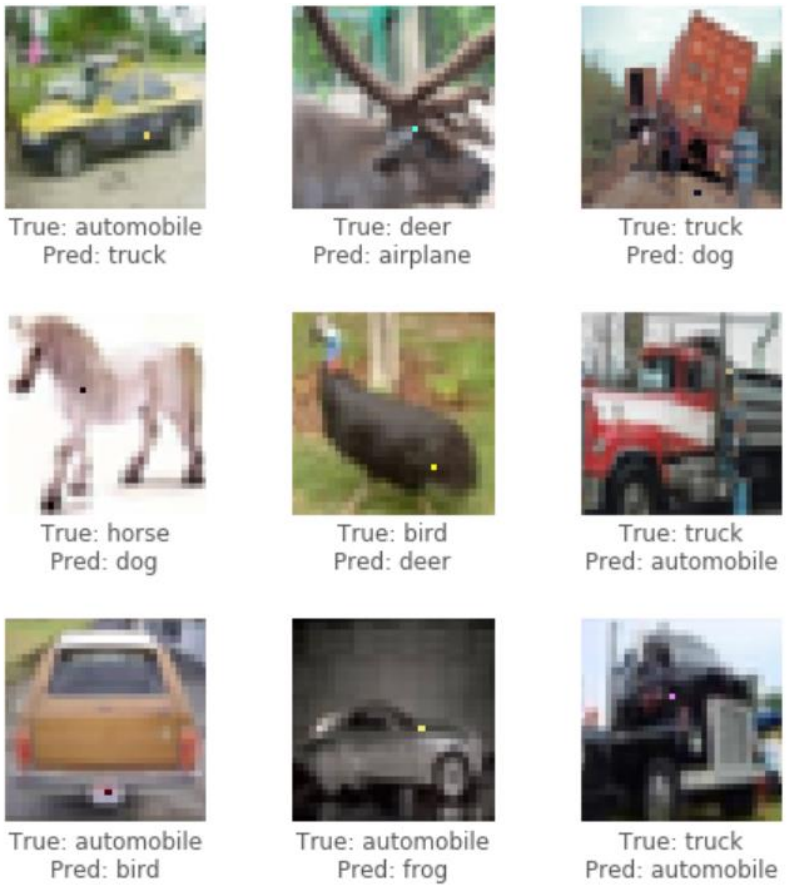
Adversarial Patch



Explaining and Harnessing Adversarial Examples



One Pixel Attack



- The paper '[One pixel attack for fooling deep neural networks]' was able to show that in some scenarios, only one pixel modification was needed to fool the classifier

Ethics

- Deep learning has the potential to be used maliciously, including in committing (cyber) crimes, subverting physical safety, or impacting negatively on geopolitical environments
- There are also dangers in using deep learning for surveillance, intelligence and military applications
- These issues are discussed in depth in many recent reports, including:
  - ‘The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation’ by a consortium of AI researchers from OpenAI and various institutions and universities across the world
  - ‘AI in the UK: ready, willing and able?’ published this week by the UK Government’s Artificial Intelligence Committee, including consultation with Google DeepMind

DeepFakes

Biases and Prejudice

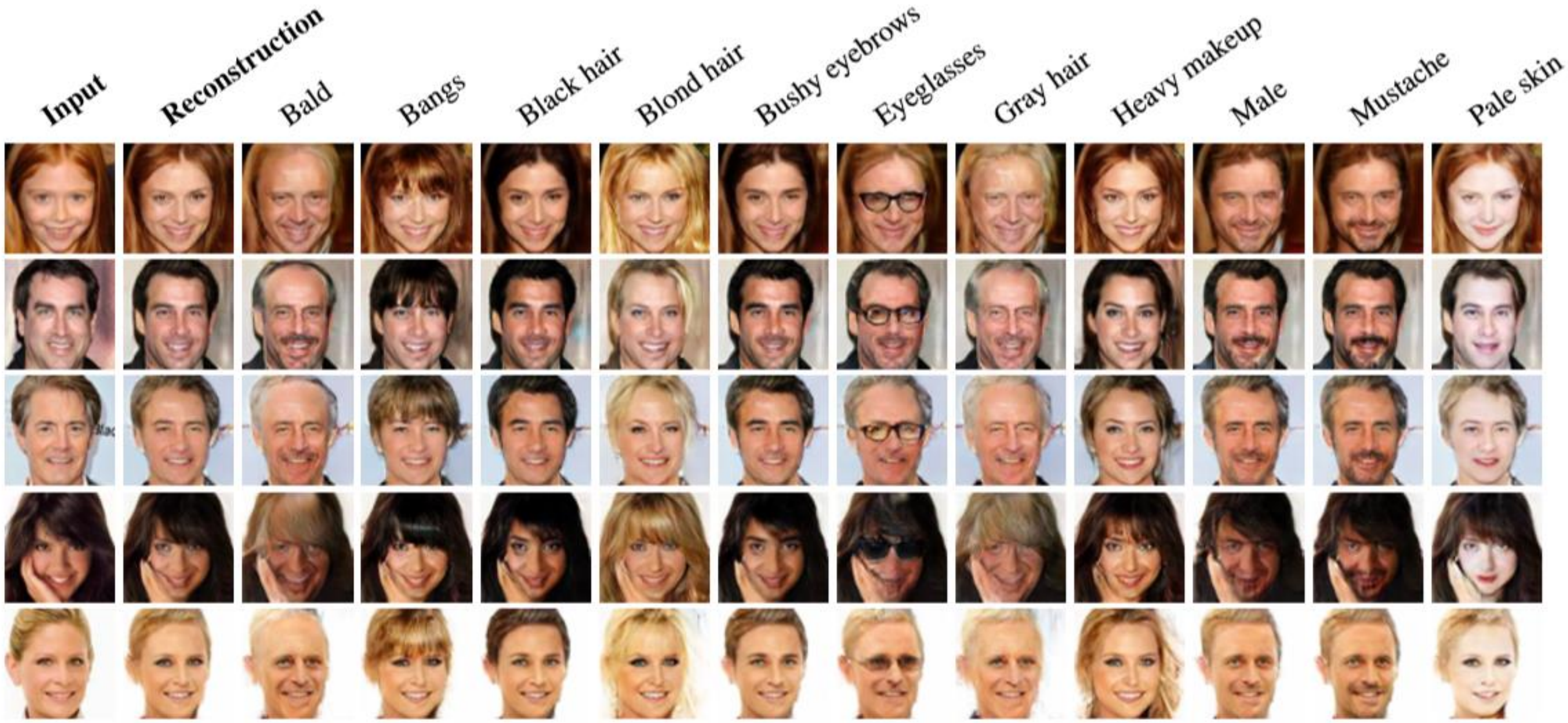
- Often deep learning results can suffer from biases or prejudices, producing outcomes that are detrimental to social liberties

- These problems may also be as a result from the dataset itself, and the way in which it was collected or annotated by human annotators
    - A model learning from our data will learn our biases

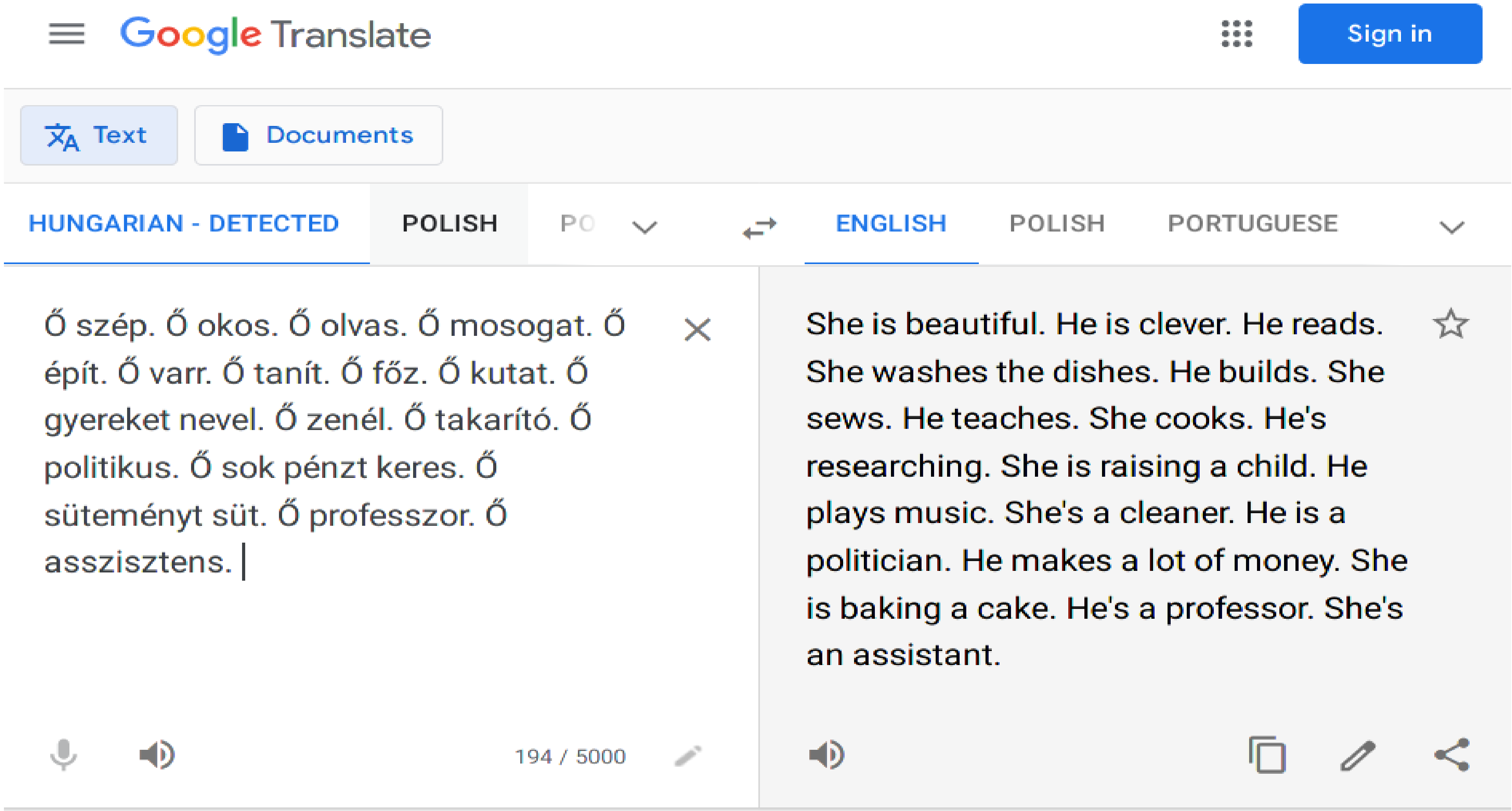
- Inversely, the development of deep learning models can actually be beneficial for exposing stereotypes and biases inherent in society, yet care must be taken to not use the models to reinforce them

Stereotypes

[Autoencoding beyond pixels using a learned similarity metric]



Language Translation



Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens.

She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant.

Predicting Sexual Orientation

- Controversy following the publishment of ‘Deep neural networks are more accurate than humans at detecting sexual orientation from facial images’ in 2017.
- The authors trained and tested their “sexual orientation detector” using 35,326 images from public profiles on a US dating website.
- However, an independent review by AI researchers showed results are due to biases in image quality, expression and grooming – i.e.. the algorithm was detecting variations based on how the users chose to portray themselves and their cultural choices, rather than their actual sexual orientation
- Automated Inference on Criminality using Face Images created similar controversy

Practical Issues

Practical Issues

- Social issues aside, there are still many barriers that AI researchers and practitioners face when developing and deploying deep learning models

- Some practical issues are actually caused as a result of social considerations, such as determining the trade-off between performance and interpretability, for example

Reproducibility

- Reproducibility and robustness of reported results is becoming much more of a focus.
- This is the reproducibility checklist implemented for the first time for submissions to the Neural Information Processing Systems conference last year. One of the top deep learning conferences in the world.

https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

Model Tuning and Assessment

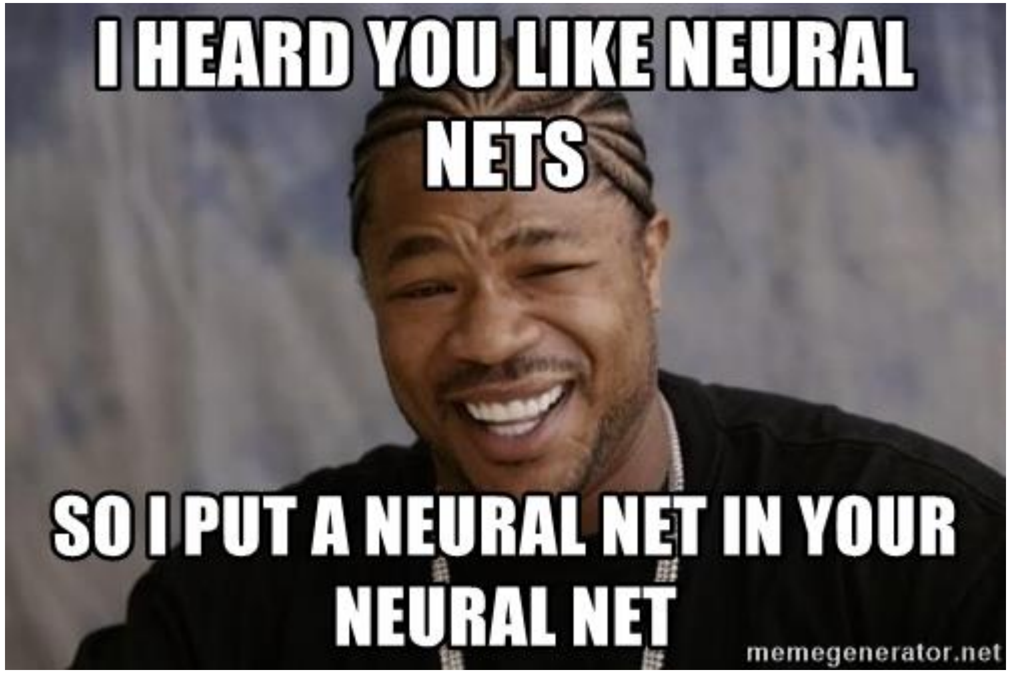
- What is the right number of features to detect? And in how many layers? What model architecture is best suited for the task? What activation functions should I use and what learning rate is best?

- Just as with feature engineering and selection, there aren’t always clear guidelines on what makes a good network architecture

- Some ways to approach this are tedious, such as trial-and-error or massive grid searches

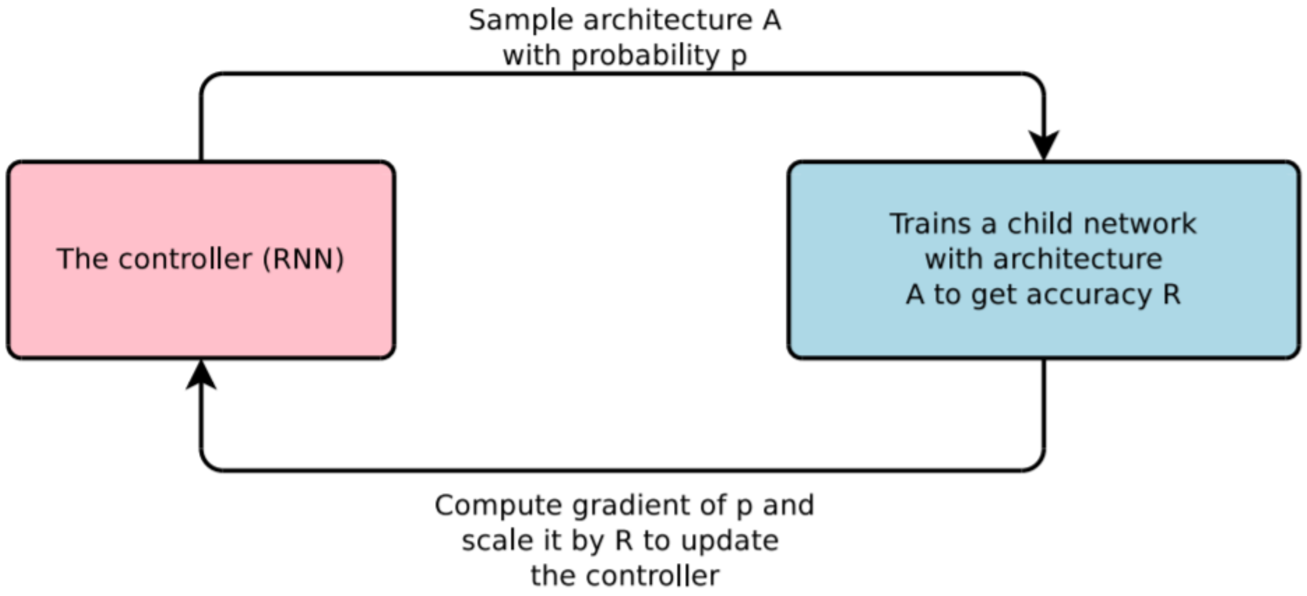
Neural Architecture Search

- Using neural networks to automate the design of neural networks – "Learning how to learn"


I HEARD YOU LIKE NEURAL NETS
SO I PUT A NEURAL NET IN YOUR NEURAL NET
memegenerator.net

Neural Architecture Search

- In ‘Neural Architecture Search with Reinforcement Learning’, the researchers employ an RNN (LSTM) to choose hyperparameters to maximise accuracies of CNN architectures on CIFAR-10 (an image classification dataset), achieving results similar to state-of-the-art hand-crafted architectures



Sample architecture A with probability p

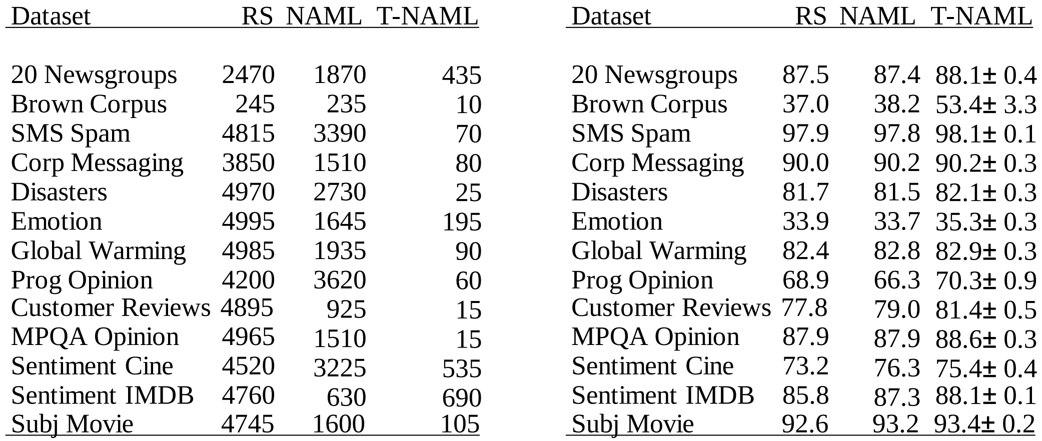
The controller (RNN)

Trains a child network with architecture A to get accuracy R

Compute gradient of p and scale it by R to update the controller

Neural Architecture Search

Multitask/Meta-learning for NAS – learning to learn how to learn



Dataset	RS	NAML	T-NAML
20 Newsgroups	2470	1870	435
Brown Corpus	245	235	10
SMS Spam	4815	3390	70
Corp Messaging	3850	1510	80
Disasters	4970	2730	25
Emotion	4995	1645	195
Global Warming	4985	1935	90
Prog Opinion	4200	3620	60
Customer Reviews	4895	925	15
MPQA Opinion	4965	1510	15
Sentiment Cine	4520	3225	535
Sentiment IMDB	4760	630	690
Subj Movie	4745	1600	105

Dataset	RS	NAML	T-NAML
20 Newsgroups	87.5	87.4	88.1± 0.4
Brown Corpus	37.0	38.2	53.4± 3.3
SMS Spam	97.9	97.8	98.1± 0.1
Corp Messaging	90.0	90.2	90.2± 0.3
Disasters	81.7	81.5	82.1± 0.3
Emotion	33.9	33.7	35.3± 0.3
Global Warming	82.4	82.8	82.9± 0.3
Prog Opinion	68.9	66.3	70.3± 0.9
Customer Reviews	77.8	79.0	81.4± 0.5
MPQA Opinion	87.9	87.9	88.6± 0.3
Sentiment Cine	73.2	76.3	75.4± 0.4
Sentiment IMDB	85.8	87.3	88.1± 0.1
Subj Movie	92.6	93.2	93.4± 0.2

Table 1: Performance of Random Search (RS), single-task Neural AutoML (NAML) and Transfer Neural AutoML (T-NAML). Left: Number of trials needed to attain a validation accuracy-top10 equal to the best achieved by Random Search with 5000 trials (250/2500 for Brown and 20 Newsgroups, respectively). Right: Test accuracy-top10 given at a fixed budget B of 500 trials (B = 250 for Brown). Error bars show ±2 s.e.m. computed across the top 10 models. Similar s.e.m. values are observed for all methods.

Neuroevolution

- The use of evolutionary computing techniques can also help to identify candidates for high performing model architectures

- For example, Uber released a suite of papers using genetic algorithms to train deep learning models

- Their results showed the potential for neuroevolutionary approaches to enhance the training of deep learning models to achieve faster convergence to higher performances

- Some of these approaches are discussed later in the course

Model Assessment

- Is accuracy always the best way of evaluating a deep learning model? (usually not for unbalanced datasets e.g. medical diagnosis 99% don’t have the diagnosis 1% do can get 99% accuracy with a model that does nothing)
- Sometimes trade-offs are necessary between other aspects of the deep learning model, such as sample complexity, model size, training times, simplicity and interpretability
- Developing large deep learning models that require hours to train on GPUs may not be worth a very small performance gain
- How does it compare to human performance?

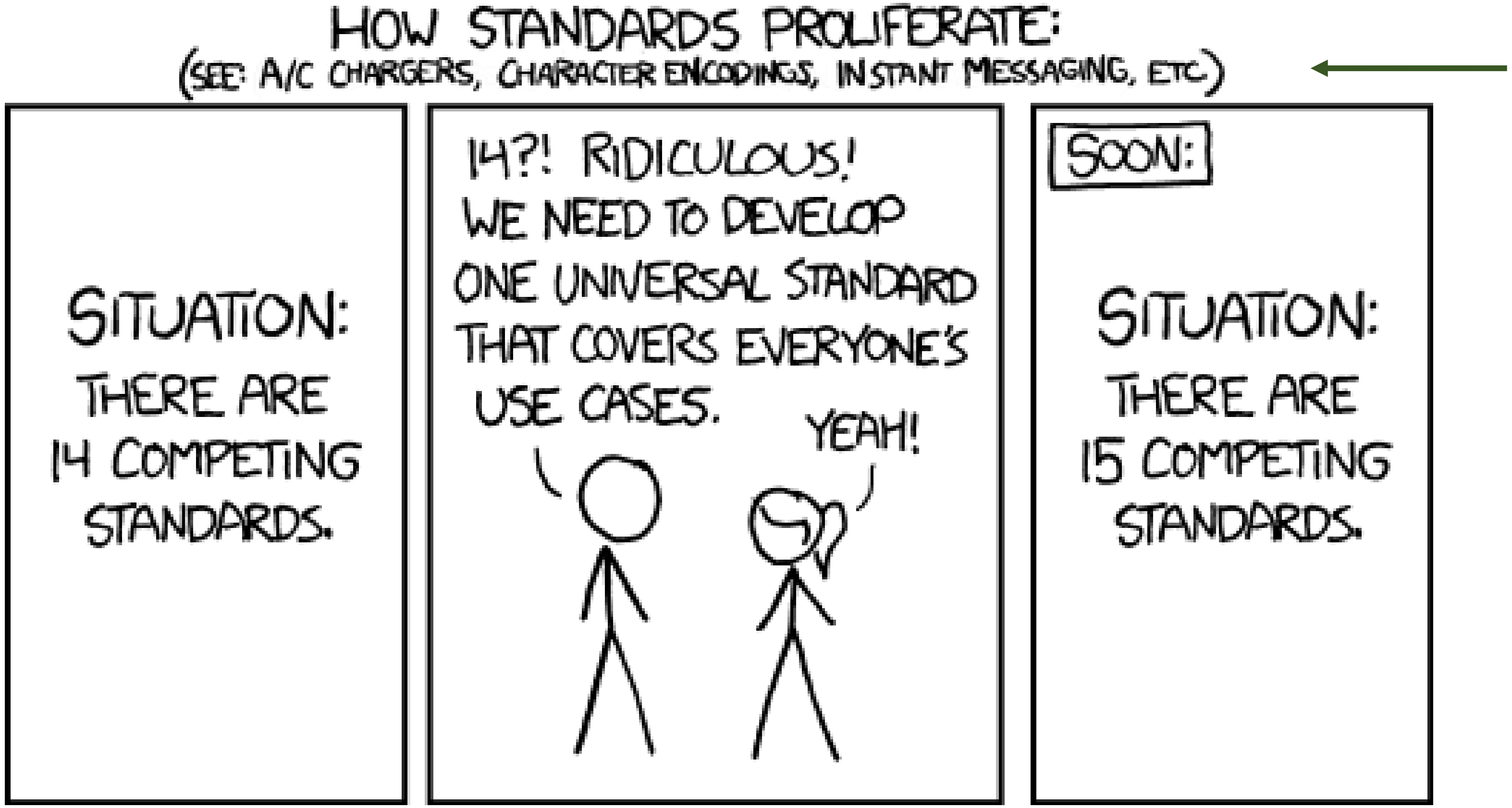
Efficiency and Practicality

- The human brain is quite efficient and able to process complex situations in very short amount of times
- Even with GPUs, models take time to train and to make inferences, particularly if they have millions of parameters
- General computing trends have seen mobile computing grow, therefore there is a need for small, high performing models capable of running on mobile phones and embedded devices
- Environmental impact considerations: Energy and Policy Considerations for Deep Learning in NLP

Small Datasets

- Humans are adaptable and capable of making reasonable assumptions with very little data
- Deep learning has typically relied upon very large datasets to train on before making accurate inferences
- Research is looking into ways to achieve success even with small numbers of training examples
  - Transfer learning is one approach
  - Deep Metric Learning – a metric is used to learn a feature space where instances of the same class are close together and different classes are further apart.
  - Few-Shot Learning – a variety of methods that focus on problems with an extremely small number of labelled training instances from each class – usually less than 20. Methods include meta-learning, prototypical networks etc.

Translatable and Transferable models


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

…and deep learning frameworks

Translatable and Transferable models

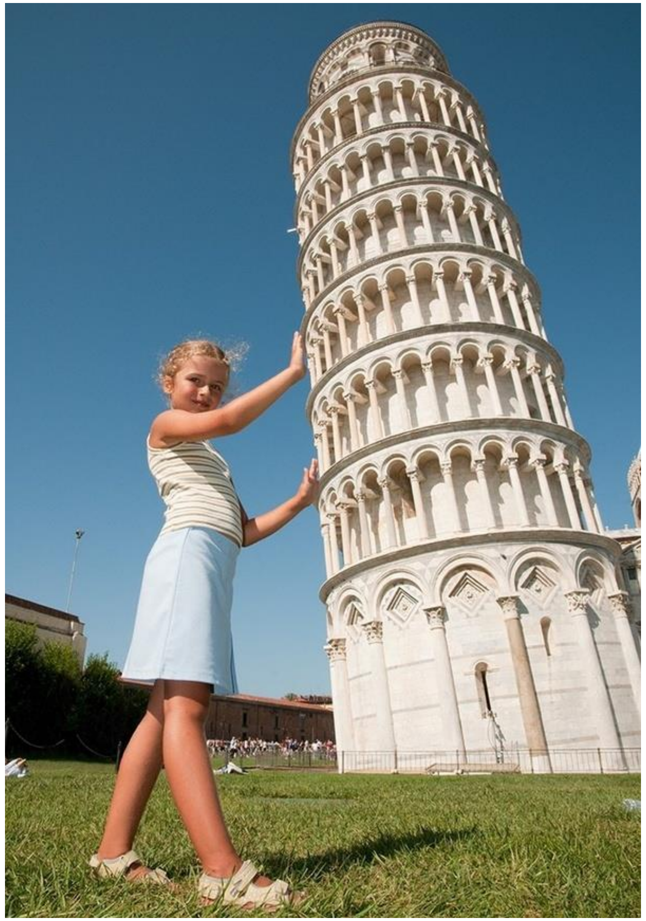
- Dozens of competing frameworks means different requirements or output formats
- The development of APIs and single reference formats aim to allow practitioners to more easily transfer models between frameworks, increasing replicability

- Gluon is an API to define deep learning models, developed by Amazon and Microsoft
- Keras provides a single interface to Theano, Tensorflow and Microsoft Cognitive Toolkit
- ONNX provides definitions for interoperable models and operators, developed by Facebook and Microsoft

Limitations Compared to the Brain

- Keep in mind that the human brain has evolved over a long time, and has specialist structures that have already been optimised prior to any formal learning taking place

- For example, can a deep learning model understand depth? Is this a tiny tower, or a giant child?



Future Directions

- Continued exploration of architecture and parameter space
- Combination of multiple approaches, including using biomimicry
- The use of deep learning and AI more generally as a data augmentation tool – providing human analysts with knowledge and advice from which they can make decisions
- Human centred machine learning
  - Brain Computer Interfaces: https://www.scientificamerican.com/article/facebook-launches-moon-shot-effort-to-decode-speech-direct-from-the-brain/

Practical Considerations

- Understand your data; it is often better to start small with simpler approaches before diving in to highly complex deep learning models
  - Be aware of potential biases in your data

- Consider the deployment of the model, and any limitations that may introduce

- Remember that the human brain has its own flaws, therefore building models that utilise the same premises can introduce problems. Likewise, a single network shouldn’t be expected to always outperform humans.

Summary

- Just like the brain, deep learning models are susceptible to ‘optical illusions’ and other attacks
- An important feature of deep learning is interpretability, and facilitating trust with users
- High complexity is not always best, as large, complicated models can be difficult to explain or deploy
- There is still much work to be done on developing clearer guidelines on model architecture development, training and optimisation, and ethical application