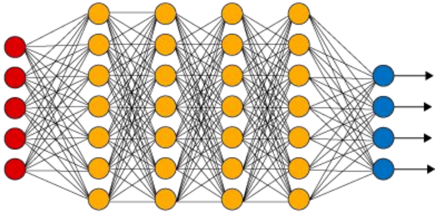
Google

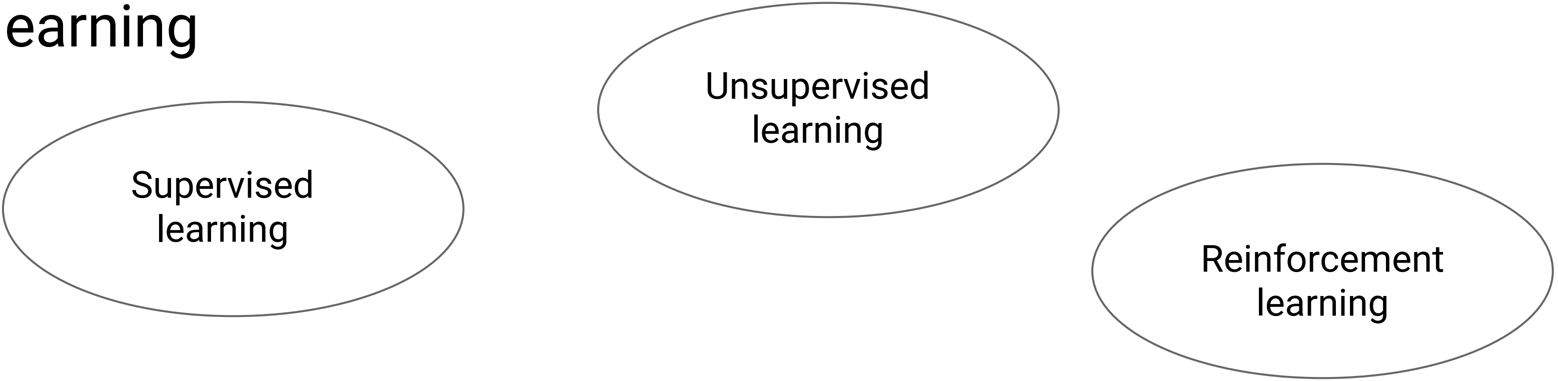


An Introduction to Deep Learning Research

Yasaman Bahri
Google Brain

KITP Physics Teachers' Conference
Automating Insight: Pushing the Frontier of Quantum Physics with Machine Learning
Feb 16, 2019

Types of learning



Supervised learning

Unsupervised learning

Reinforcement learning

Machine learning

- Draws from: computer science, statistics, ….

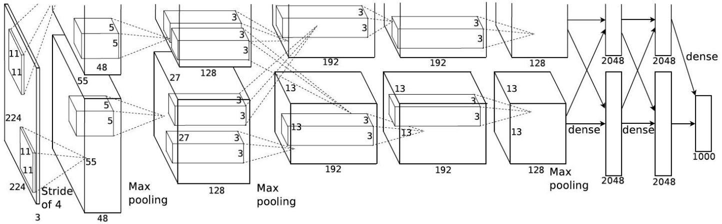
Artificial Intelligence

- Draws from: computer science, statistics, cognitive science, neuroscience, linguistics, ….

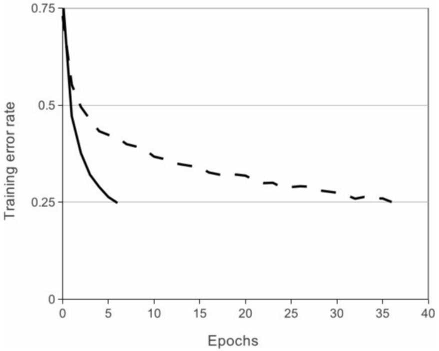
Why is deep learning so popular now?

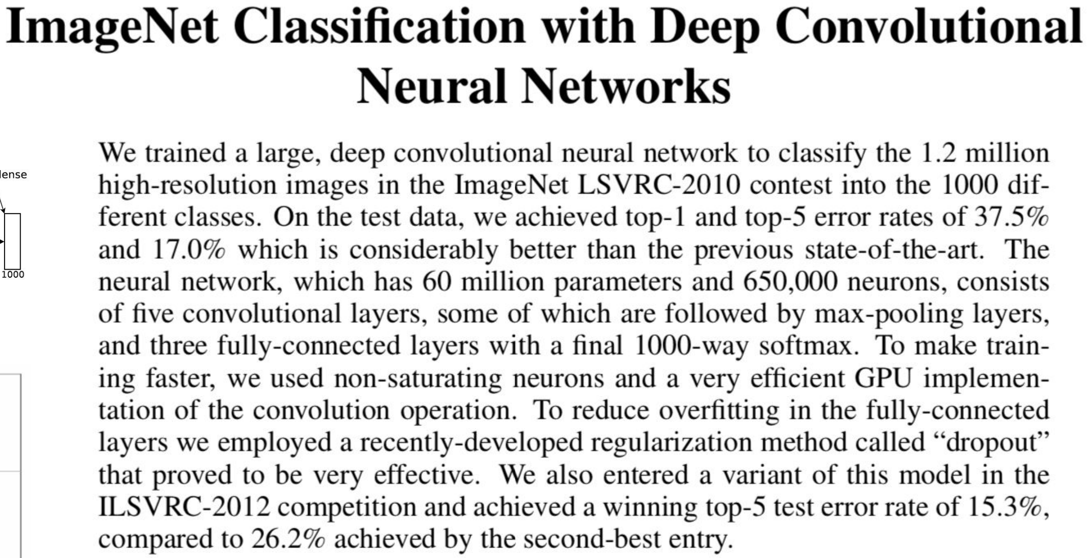
Convolutional neural networks -- from the 90’s (Y. LeCun, et al.)

(“AlexNet”)



Changing nonlinearity allows to train faster (6x)





ImageNet Classification with Deep Convolutional Neural Networks

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

A. Krizhevsky, et al. NeurIPS 2012.

Elements of deep learning in the modern age (large scale)

Access to large amounts of data

- Also, at the level of research, constructing datasets/establishing tasks of greater complexity helps research dramatically

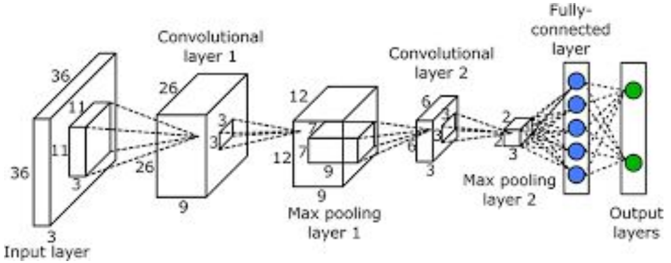
Very large (overparameterized) models

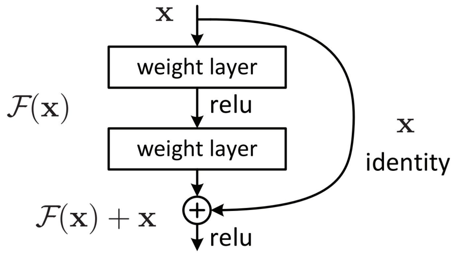
More compute

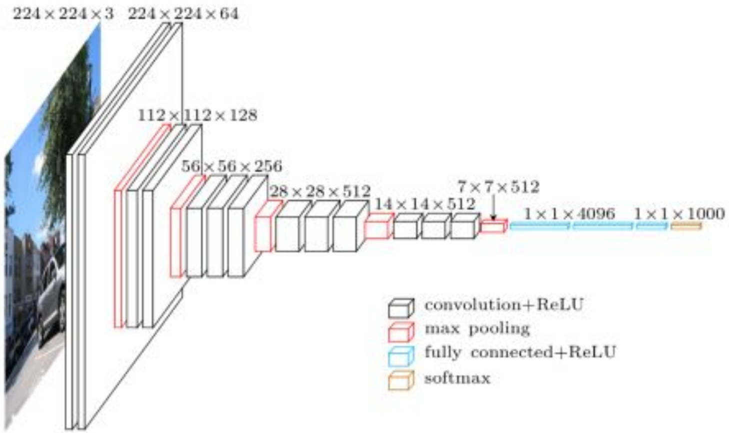
Zoo of models!

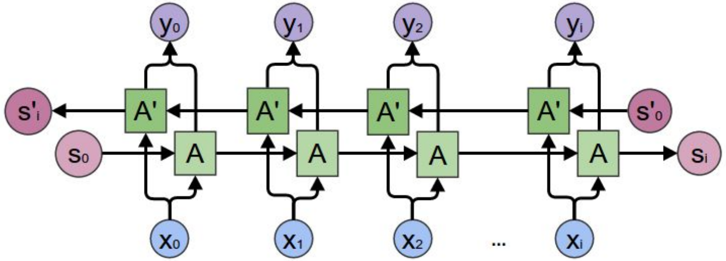
Some primitives:

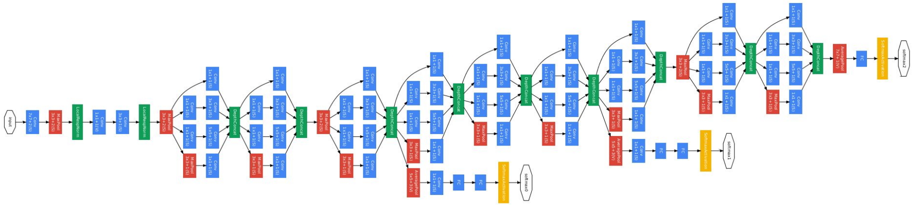
- Convolutional filters
- Pooling
- Recurrent structure
- Skip connections
- Attention, External Memory











Different ways of producing predictions

Empirical Risk Minimization

$$z(\theta, x) \rightarrow \text{ need to determine } \theta$$

Minimize loss, which is a function of the parameters of the network

$$\mathcal{L}(\theta) = \frac{1}{M} \sum_{i=1}^{M} C(z_i, y_i) \text{ where for instance } C(z_i, y_i) = (z_i - y_i)^2$$

Bayesian inference

Make predictions in a manner consistent with **Bayes rule** from probability

$$\text{Prior } p(\theta) \rightarrow \text{ Posterior } p(\theta|\mathcal{D})$$

$$p(z|\mathcal{D}, x) = \int d\theta \, p(z|\theta, x) \, p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \int d\theta \, p(z|\theta, x) \, p(\mathcal{D}|\theta) \, p(\theta)$$

Field is relatively young (compared to physics)

From the Institute for Advanced Study, Program on Theoretical Machine Learning (description)

“Design of algorithms and machines capable of ‘intelligent’ comprehension and decision making is one of the major scientific and technological challenges of this century. It is a challenge because ….

- *...it calls for new paradigms for mathematical reasoning, such as formalizing the ‘meaning’ or ‘information content’ of a piece of text or an image or scientific data.*

- *...the algorithms involved must scale to very large input sizes.*

- *...because the obvious ways of formalizing many computational tasks in machine learning (from theoretical computer science) are provably intractable in a worst-case sense, and thus calls for new modes of analysis.”*

Has machine learning become alchemy?

Are we stumbling in the dark? Reliance on trial and error.

Sometimes stripping complexity from state-of-the-art reveals that you didn't need it. Other times, performance comes from tricks on top, because core approach is flawed.

- More ablation studies to understand what different components are doing?
- Reproducibility can be an issue

> "Ben Recht, a computer scientist at the University of California, Berkeley, and coauthor [of the keynote talk], says AI needs to borrow from physics, where researchers often shrink a problem down to a smaller 'toy problem.' 'Physicists are amazing at devising simple experiments to root out explanations for phenomena,' he says." [1]

"Science of deep learning"

[1]. *https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy*
Also, Ali Rahimi's Test-of-time award talk from NeurIPS 2017 -- check it out on YouTube.

Aside: educational trajectory & current research interests

B.A. in Physics & Math; Ph.D. (‘17) in theoretical condensed matter physics (UC Berkeley)
Thesis work was in quantum many-body theory.

Started research at Google Brain after Ph.D.

Research motivations: scientific understanding of deep learning. Principled approaches.
Tools: sometimes analytic techniques from theoretical physics.

Can the “many-body” perspective & experience building theory for complicated interacting systems help with machine learning and, more generally, AI?

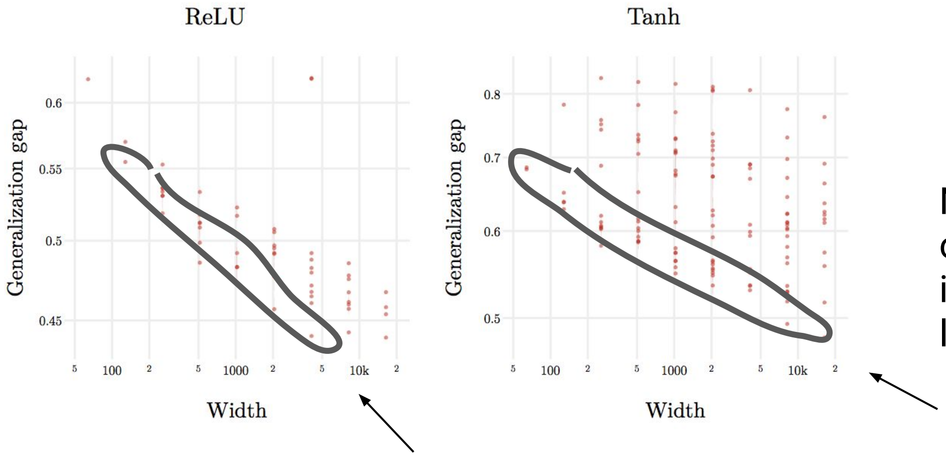
These fields are a mixture of science, engineering, and design.

Phil Anderson. “More is Different.” Science 04 Aug 1972: Vol. 177, Issue 4047, pp. 393-396.

The overparameterized regime of deep learning

Observation: why do *"large"* networks generalize well? Why aren't you plagued with overfitting when you add more parameters?

Unexplained by older, classical results in *statistical learning theory*. (Think curve fitting with polynomials!)

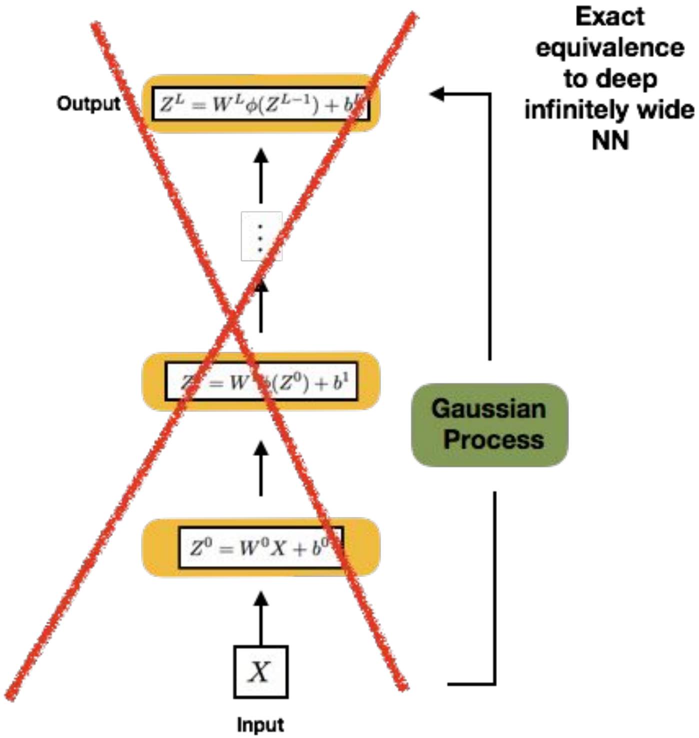


Networks trained with gradient-descent optimization: best achievable test performance improves with number of units in a hidden layer.

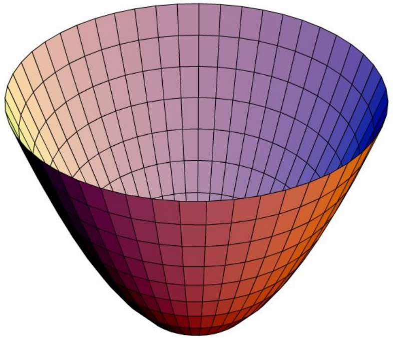
What happens in limit of infinite width?

Deep neural networks and Gaussian process correspondence

By (i) taking a particular limit and (ii) conceptually shifting our perspective on the problem, we can map one seemingly complex problem onto another problem that is easier to understand.


Output
$Z^L = W^L\phi(Z^{L-1}) + b^L$
$\vdots$
$Z^1 = W^1\phi(Z^0) + b^1$
$Z^0 = W^0X + b^0$
X
Input
Gaussian Process
Exact equivalence to deep infinitely wide NN

Optimization landscape of deep learning

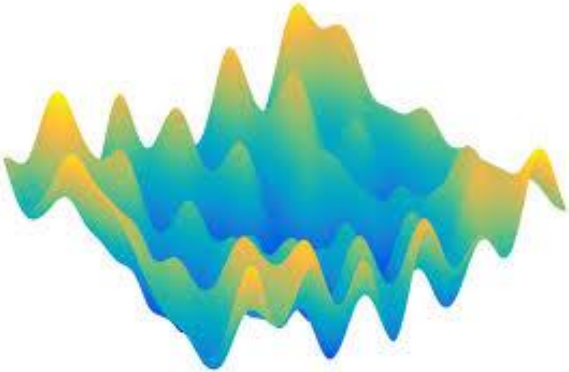


Historically, tried to have some guarantees on optimization

- Convex optimization

Now, wish to use a much richer class of (differentiable) models, but unclear how difficult the optimization problem is!

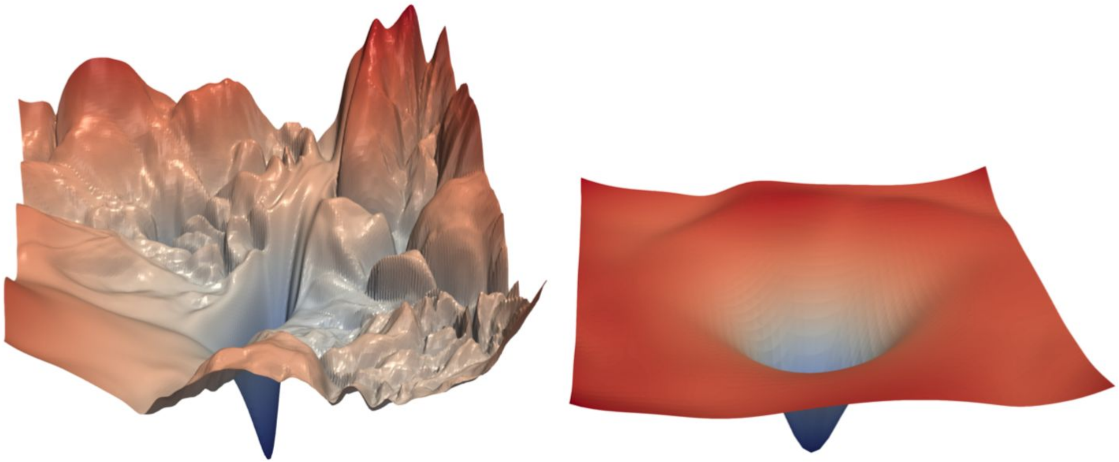
- Nonconvex optimization



Our geometric intuition can be very misleading in high-dimensional spaces!

Optimization landscape of deep learning

Tension between guaranteed training (algorithms) and a richer but less understood class of models.



(a) without skip connections (b) with skip connections

The “magic” of deep learning is partly related to the optimization process.

Image from H. Li, et al. NeurIPS 2018.

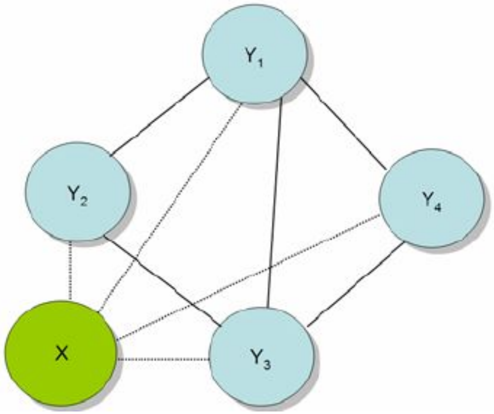
Physics and machine learning in the past

Spin glass models

Probabilistic graphical models

(from Statistical Mechanics)

Bayesian inference

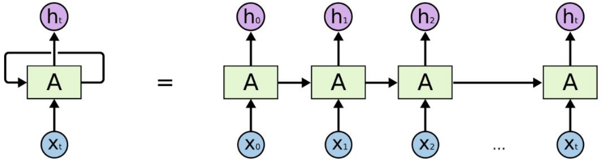


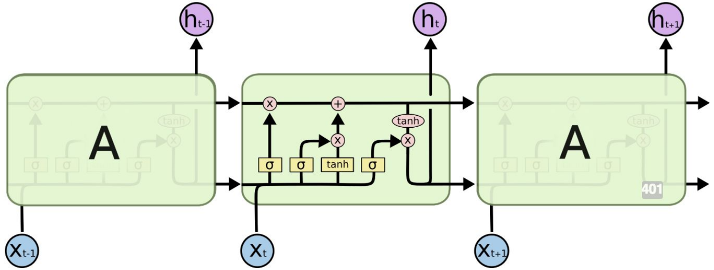
Ways of thinking about systems with many random variables.

Snippets of Current Research (Successes and Susceptibilities)

Models for sequences: new components

Recurrent neural networks (RNNs) for problems with time translation invariance (sequences)





LSTM (Hochreiter & Schmidhuber '97)

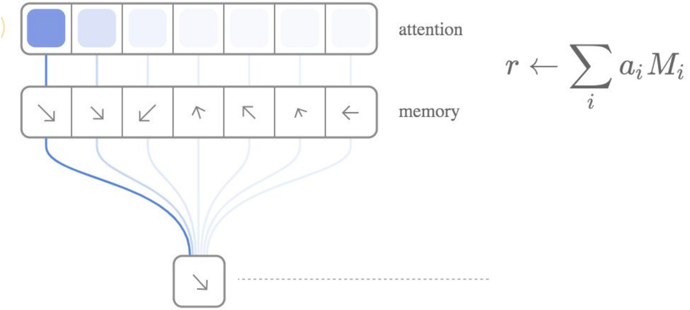
--Usage of a cell state in addition to a hidden states, able to forget information, controlled by gates (nonlinearities)

The challenge of how to deal with long-range dependencies: relevant to language, music, speech, ….

One advancement: idea of attention
--Tells e.g. an RNN cell where to look (for instance, looking at external memory if model has it or looking at another sequence)

Visuals: credit to Chris Olah

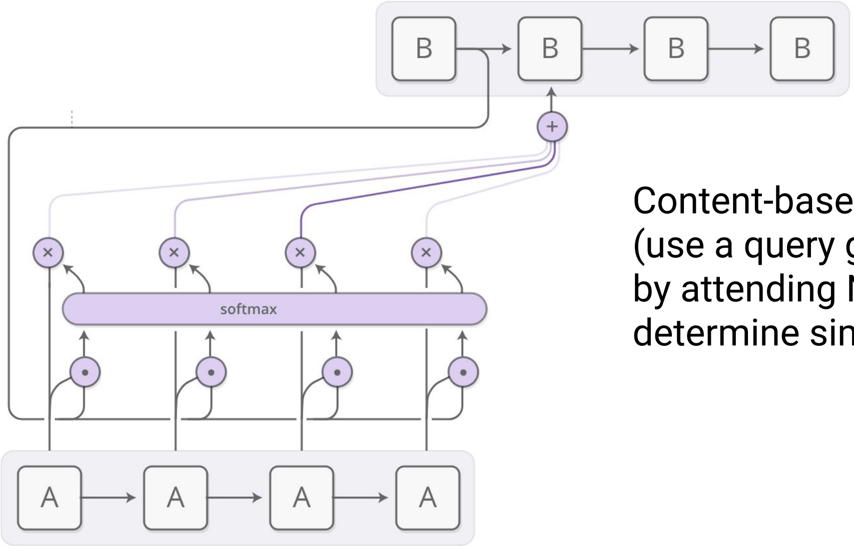
Models for sequences: new components



Keep network differentiable -- “magic” of gradient-based optimization! “Differentiable programming”

Focus with different weights (attention weights) at all locations.

Reading from memory using attention (See e.g. “Neural Turing Machine” as an early model that used this)



Content-based attention (use a query generated by attending NN to determine similarity)

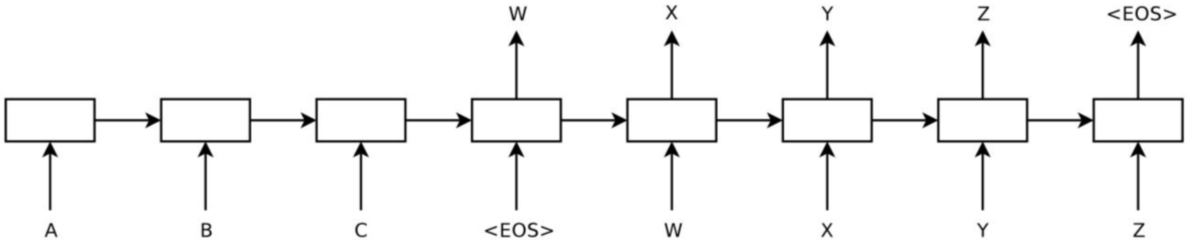
Visuals: credit to Chris Olah

Google

Neural machine translation (NMT)

Traditional sequence-to-sequence model:

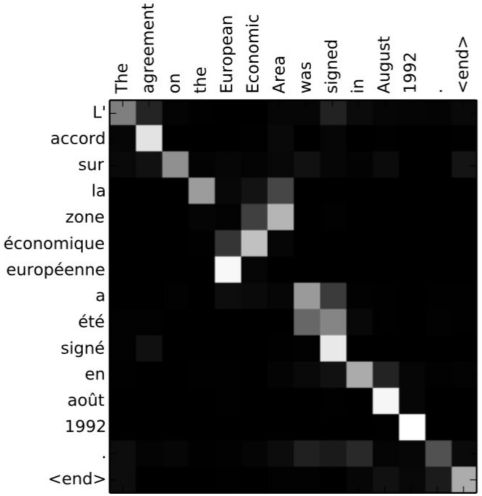
- Involves an encoder-decoder where encoder compresses source sequence into a fixed length vector (context vector). Decoder predicts next word in translation given the context vector and all the previous generated words in translation.

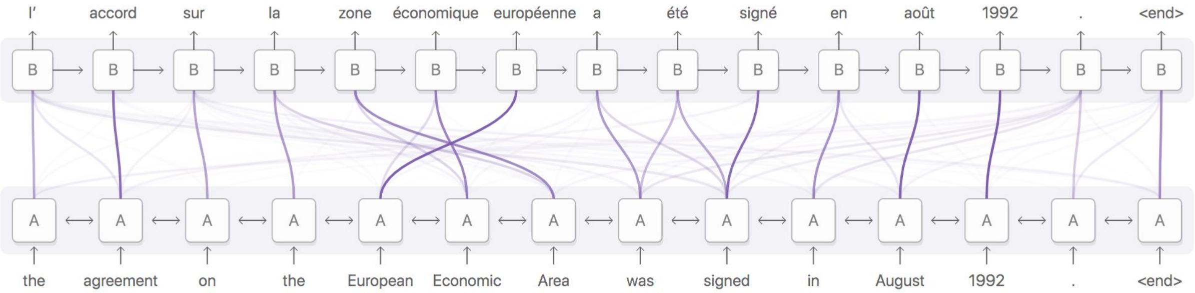


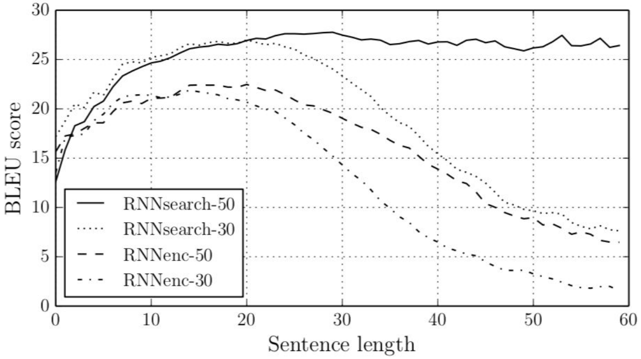
From Sutskever, et al. NIPS 2014.
Also check out Tensorflow tutorial on Neural Machine Translation: https://github.com/tensorflow/nmt

Neural machine translation (NMT)

Can use attention to go beyond this (e.g. “jointly learning to align and translate”)







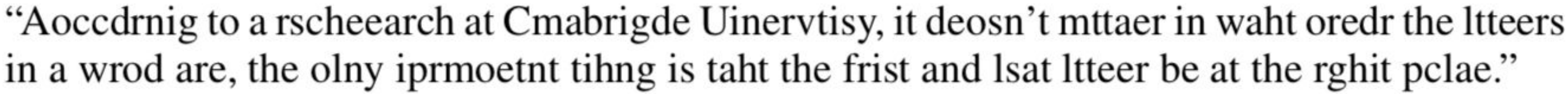
“Jointly learning to align and translate”
Bahdanau, et al. ICLR 2015.

Example failure modes and susceptibilities

Many recent success stories in language as a result of deep learning

- Translation, question/answering, text generation….

On the other hand, systems can exhibit fragility. Examples of this include susceptibility to even a little bit of noise (random, swapping, or human error) and also hallucinations (insertion of one token that wreaks havoc!).



“Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn’t mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae.”

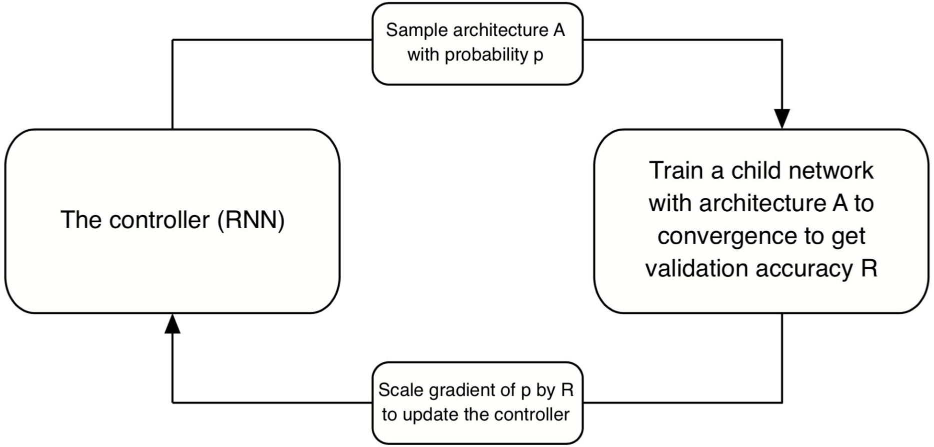
Are networks just memorizing? How far can we go with statistical correlations? How far can one go with the “differentiable programming” paradigm in language understanding?

[1]. See Belinkov, et al. ICLR 2018.

AutoML: neural architecture search

Significant architecture engineering currently involved: can we automate the process of machine learning?

“Neural Architecture Search”: can we search intelligently over the space of models?



Sample architecture A with probability p

Train a child network with architecture A to convergence to get validation accuracy R

Scale gradient of p by R to update the controller

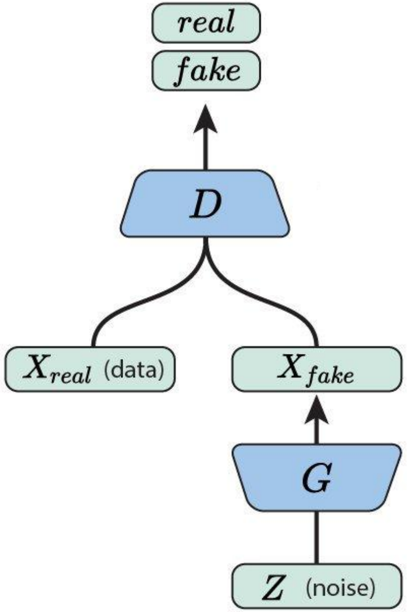
The controller (RNN)

Trained on one dataset to produce architecture which performs well on another dataset that is similar (images)

B. Zoph, et al. CVPR 2018.

Generative models for images

Can we generate samples from some probability distribution of interest? Idea of a ***generative model***.



I. Goodfellow, et al. NeurIPS 2014.

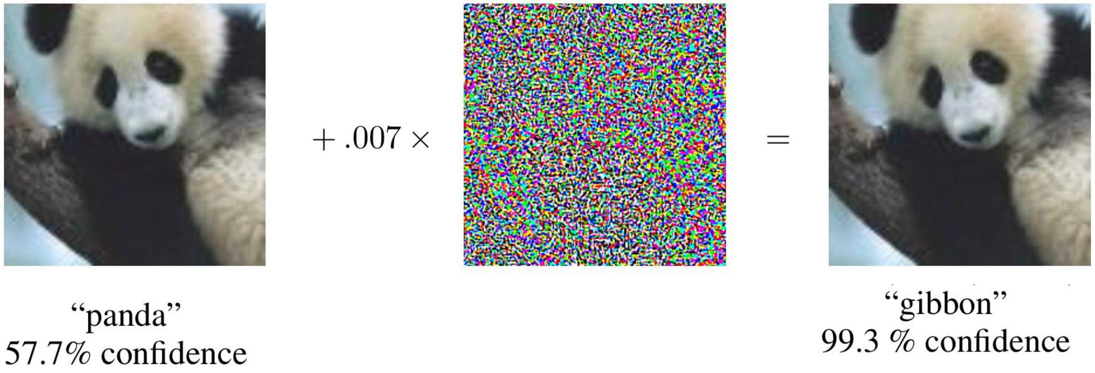
Generative Adversarial Networks are one particular realization of this.

- Use ideas from **game theory**: consider a two-player game between a **generator G** and a **discriminator D**
- The entire system is trained ***end-to-end***

Progressive GANs: T. Karras, et al. ICLR 2018

Is it sampling from the true underlying distribution? Is it “memorizing” or “learning”?

Lack of robustness: adversarial examples



Solutions considered:

- Adversarial training (training on such examples)
- Can we design architectures differently?

Theoretically, one difficulty is we that don't know how to characterize the "data manifold." So return to toy problems.

These examples can be found efficiently -- but are particular directions in input space, not chosen randomly.

In high-dimensional spaces, will any model necessarily have them?

On the other hand, we are prone to errors too -- it's just that the types of errors we are susceptible to are different from what neural networks are susceptible to.

C. Szegedy, et al. ICLR 2014. I. Goodfellow, et al. ICLR 2015.

Privacy & Security

- Performance isn’t all that matters in the real world: additional design constraints like privacy → ***“ML in the wild”***

- White box (access to model parameters) vs black box attack (no access, just queries)

- Despite not having access to training data (user data), may be able to infer some aspects of the data through specialized attacks!

- Theoretical frameworks have been developed to think about these tradeoffs between performance and privacy.

Interpretability of Neural Networks

Interpretable machine learning: for many use cases, would like neural networks to be able to “explain” how they reached a decision

Feature visualization and attribution

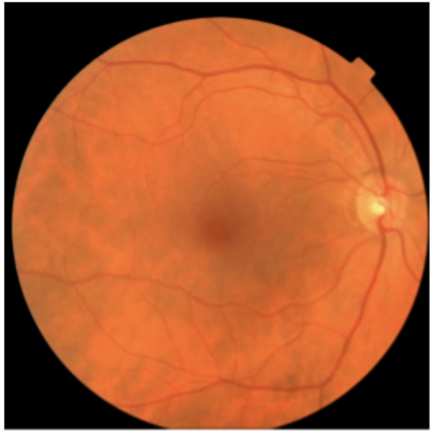
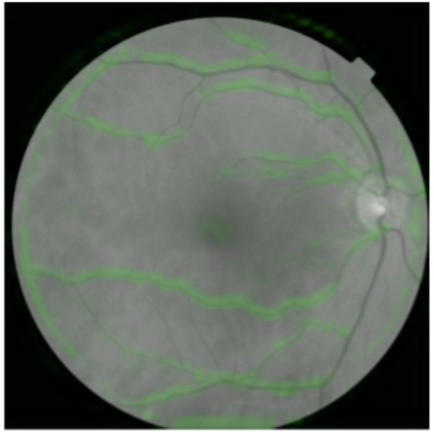


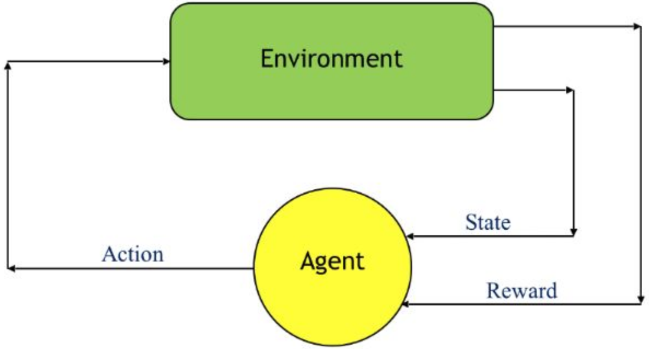
Image of retina



Blood pressure predictions focus on blood vessels

https://ai.googleblog.com/2018/02/assessing-cardiovascular-risk-factors.html

Reinforcement Learning (RL)



One of the most exciting frontiers!

Framework of RL: a Markov decision process. Characterized by {States, Actions, Probabilities, Rewards}.

Want to learn a function π: states (s) → action (a)

Probabilities P(s' | s, a) which determine transitions

Rewards R(s, a, s') that you obtain

Big successes in playing games (AlphaGo, AlphaGo Zero from DeepMind).

- In particular, AlphaGo Zero trained without the need for human games (learning through self-play).

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last – a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
SONGBIRDS À LA CARTE
SAFEGUARD TRANSPARENCY
WHEN GENES GOT 'SELFISH'

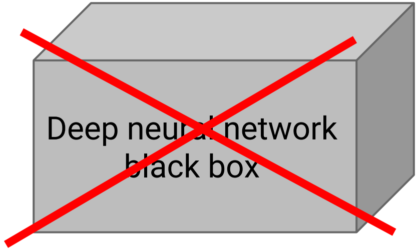
Just a few of the challenges in this domain:

- Sparsity of rewards
- Exploration vs exploitation tradeoff

Closing

Nearer Term

Principled architecture design and algorithmic choices in deep learning
Theoretical understanding, foundational concepts and definitions
Reliability, security, interpretability for usage in the real world
…..



Longer term

“Solving” a domain (e.g. vision, language, robotics), …..
Putting pieces together to get AGI (artificial general intelligence). Unclear what this even involves!

Resources for staying abreast of research developments

Google AI and OpenAI Blogs

Distill journal
https://distill.pub/

Tensorflow tutorials
https://www.tensorflow.org/tutorials

Excellent university courses

Feel free to contact me! Email: yasamanb@google.com

Thanks for your interest!

Google