6.034

Deep Neural Networks

Peter Szolovits

ai6034.mit.edu

October 9, 2019

How Does Human/Animal Vision Work?

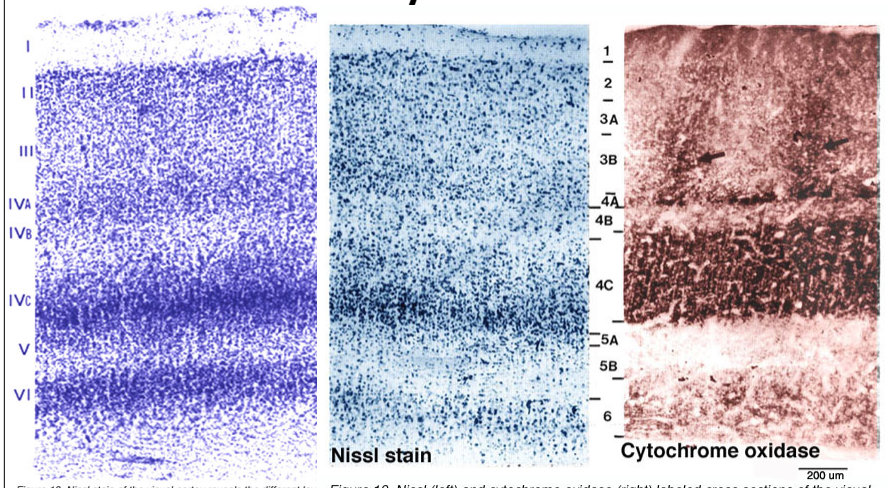


Figure 13. Nissl stain of the visual cortex reveals the different lay quite clearly.

Figure 10. Nissl (left) and cytochrome oxidase (right) labeled cross sections of the visual cortex of a macaque monkey, showing the individual layers.

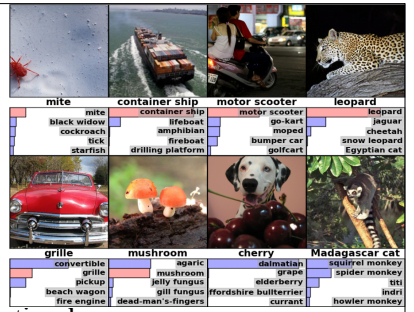
https://webvision.med.utah.edu/book/part-ix-brain-visual-areas/the-primary-visual-cortex/

Listen in on Neuroscientists Speak

- **"Layer 1 is composed of a dense network of synapses formed between the apical dendrites of layer 2-5B pyramidal cells (Lund and Wu, 1997) and the collected inputs from LGNd K layers (Fitzpatrick et al., 1983; Lachica and Casagrande, 1992), the pulvinar, feedback pathways from extrastriate areas, nonspecific thalamic nuclei, and other subcortical regions. Thus, while layer 1 contains few neurons, it is a networking layer that has a direct concerted affect on the firing properties of pyramidal cells in deeper layers."**
- **"Many types of columns have been proposed including ocular dominance, orientation, spatial frequency, and color columns."**

https://webvision.med.utah.edu/book/part-ix-brain-visual-areas/the-primary-visual-cortex/

ANN: How big?

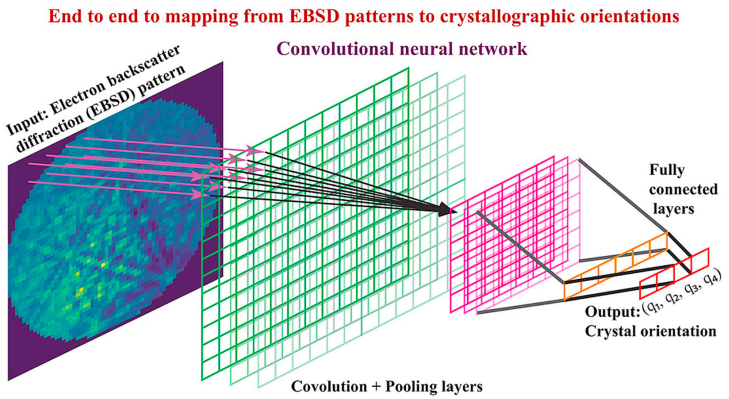


- Hinton's network
  - 650,000 neurons
  - 60 million parameters
  - 1000 neurons in final classification layer
  - 1.2 million training examples
  - Five to six days on pair of big GPUs
  - Top 1 error rate: 37%, Top 5: 17%
  - 8 layers
- Today's GPT-2 network: ~1.5B parameters
- Various architectures

How Does a Convolutional Net Work?

- (Typically) 2-D matrix of pixel values
- Kernels compute some linear combination of values in a sub-matrix, then apply non-linearity
  - Stride determines if we reduce dimensionality of image
- Pooling layers reduce dimensionality and summarize sub-matrix data
- Use multiple kernels
- Feed-forward (possibly multi-layer) neural net computes results

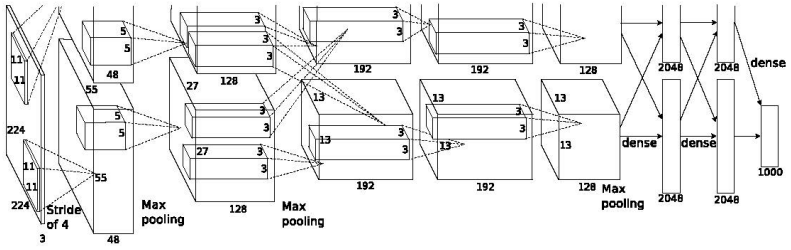
Diagram of a Simple CNN



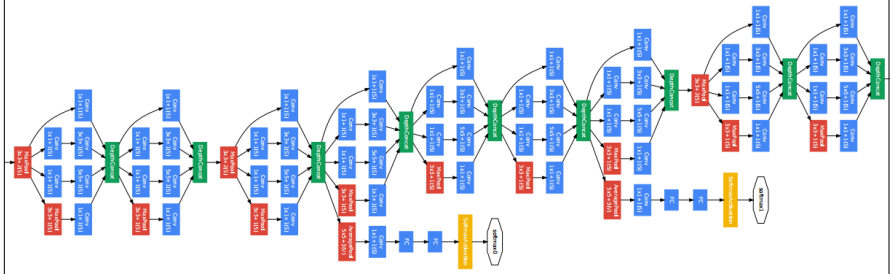
https://ars.els-cdn.com/content/image/1-s2.0-S1359645419301697-fx1_lrg.jpg

Hinton's ImageNet CNN

- *Additional Tricks:*
  - Brightness normalization
  - Translation and reflection of training data
- PCA representation of colors
- Dropout
- Overlapping pooling regions

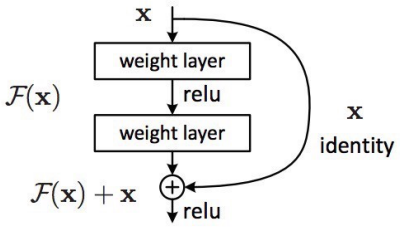


GoogleNet



https://towardsdatascience.com/an-intuitive-guide-to-deep-network-architectures-65fdc477db41

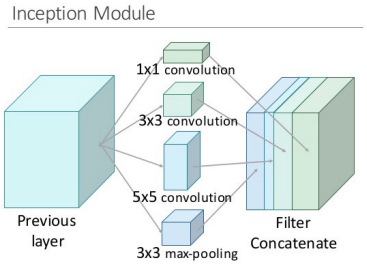
ResNet



- Each layer learns residuals
- Helps with "vanishing gradient" in very deep networks
  - Possible to train 1000+ layer networks!

https://towardsdatascience.com/an-intuitive-guide-to-deep-network-architectures-65fdc477db41

Inception



- Automate choice of "hyperparameters" by combining various filters at each layer
- "Kill 'em all and let G*d sort 'em out"
- Slooooow…

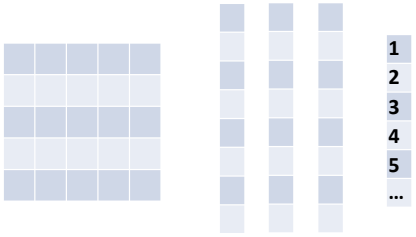
⇒Tricks to compress data

- e.g., 1 x 1 x *n* convolutions to shrink previous layers

https://towardsdatascience.com/an-intuitive-guide-to-deep-network-architectures-65fdc477db41

Compare to Pre-Massive-Computation Era

- MNIST data set of hand-written digits
- 28x28 grayscale pixels, interpolated from 20x20 B&W images
- 60K training, 10K test

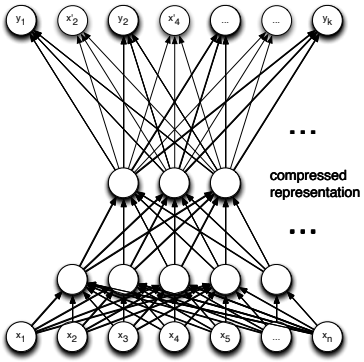


Autocoding Using Deep Neural Networks

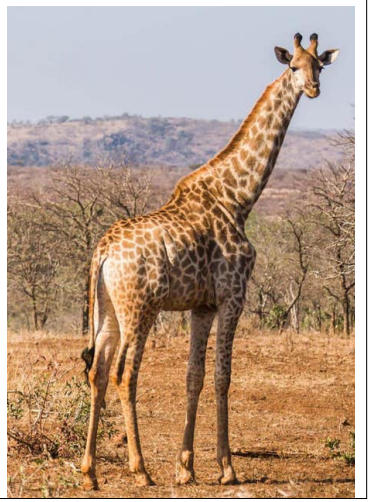
CSAIL

- Every node gets a logistic regression function of its inputs
- Number of nodes in each layer may vary
- Number of layers is another hyper-parameter
- Dropout may omit some fraction of links
- Training by back propagation
  - change weights in proportion to error signal

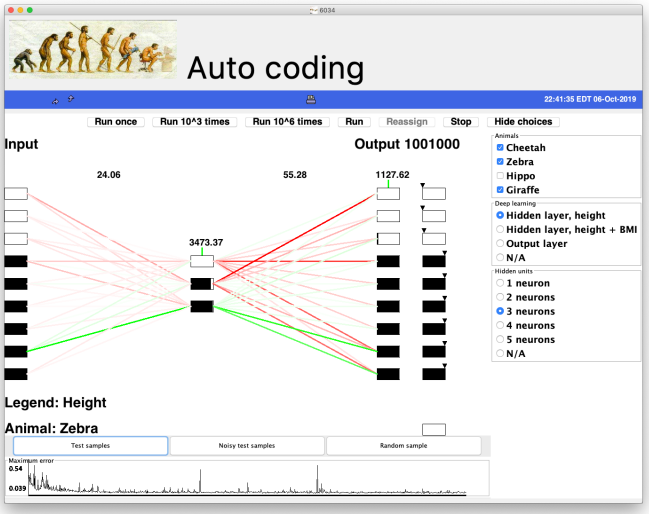
- **Unsupervised**: Train to optimize unsupervised compression
- **Supervised**: Train to objective function on gold standard data



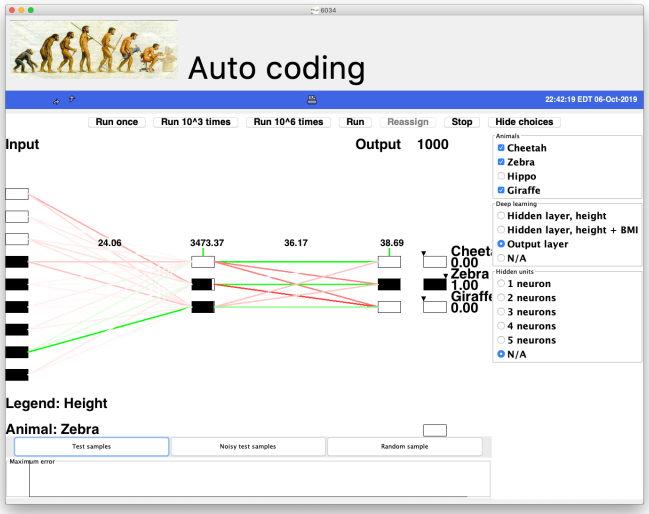
Demo: Recognize Animals from the Height of their Shadow



Autocoding for Recognizing Animals from their Shadows



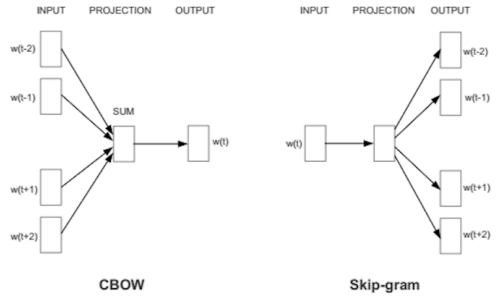
Fine-Tuning Needs Few Runs



Dropout

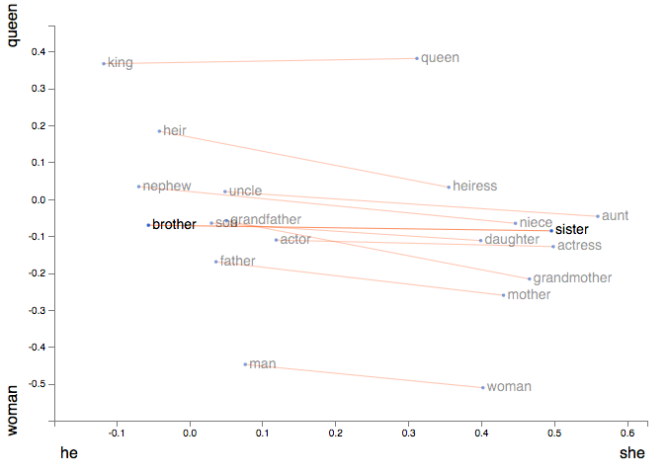
- On each training iteration, disconnect outputs of some fraction of internal neurons
- Helps prevent
  - getting stuck at local minima
  - overfitting training data

How to Turn Non-Numeric Data into Numbers



- Create a 1-layer NN to predict a word from its context or the context from a word
- Use the weights of the trained NN as a high-dimensional vector representation of the word

Vector Space Embeddings Reflect Meaning Similarity and Analogies

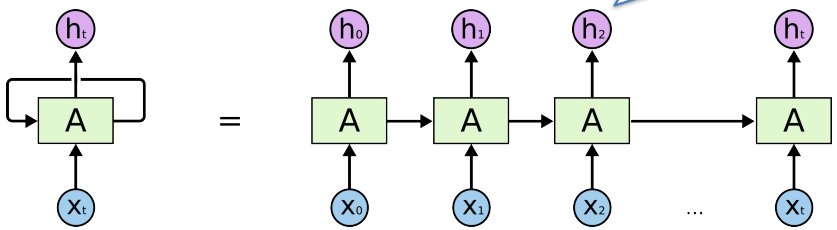


https://p.migdal.pl/2017/01/06/king-man-woman-queen-why.html

Recurrent Neural Networks

- Natural for serial data
  - Time series, e.g., measurements such as ECG, lab values
  - Natural Language text
  - Motion Pictures

Chris Olah's blog has excellent explanations



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

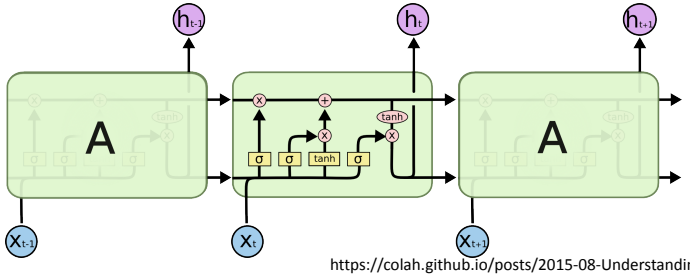
Augmentations to RNN Models

- Long Short Term Memory
- Sequence-to-sequence
- Attention

- Everything must remain differentiable!!!

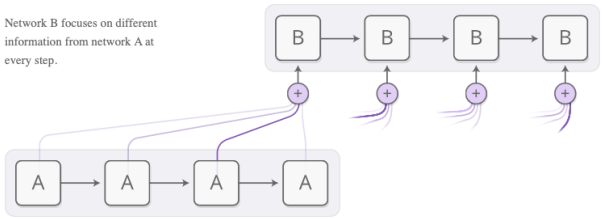
LSTM (Long Short Term Memory)

- Complicate each cell:
  - Add “Forget”, “Input”, and “Output” gates
  - Helps long-distance propagation of information
  - Many variations on this architecture also



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

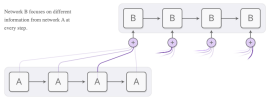
RNN Seq-Seq Model

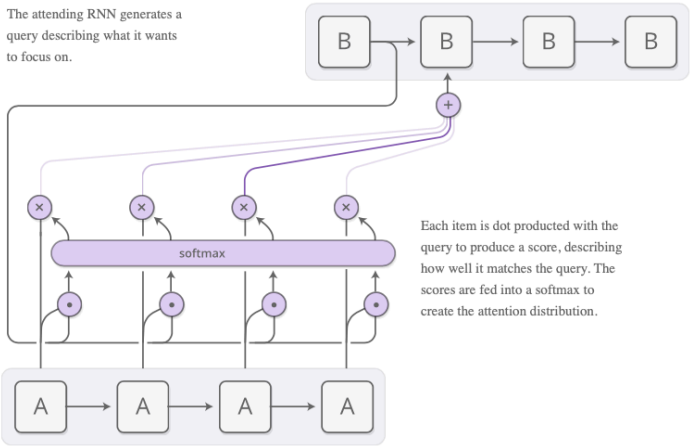


Typical sequence-to-sequence model for machine translation

https://distill.pub/2016/augmented-rnns/

Attention





https://distill.pub/2016/augmented-rnns/

Attention in Machine Translation

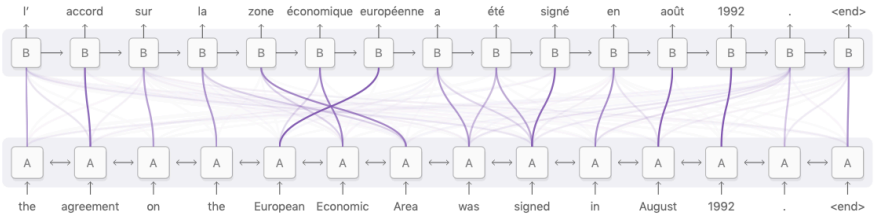


Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

https://distill.pub/2016/augmented-rnns/

Attention in Speech Transcription

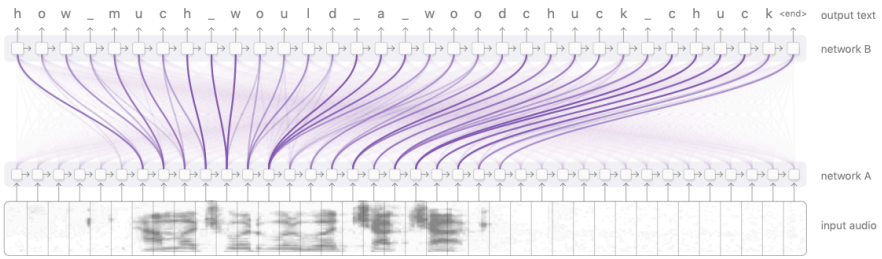


Figure derived from Chan, *et al.* 2015

https://distill.pub/2016/augmented-rnns/

Attention Helps Explanation

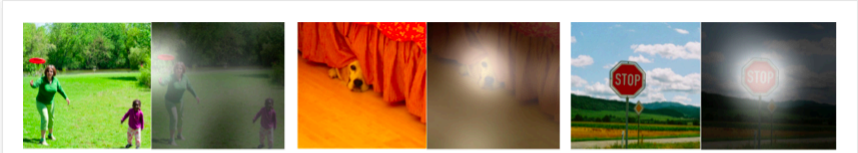


Figure from [3]

Language Models

- **Create n-tuple models from a corpus, say Shakespeare**
- **Generate new "Shakespearean" text from conditional probabilities of next character given the previous *k* characters**

k-tuple (character) Model of Shakespeare (*k*=2)

Fir lace, mes. Pat iserat dots asur do?

Dot oath my op apparm.
ABEROCTAGAR:
My me, prep day.

LANA:
PATHUMBET:
Hows. Younignionesad cork. Thas,
I'll theall no. Whanspeart,
ALBOTENTIO:
On, to theshat youl werchind, hown:
AUS:
Said sioner offe swillat unusbanould is the ofte, thistatiescall badvannow he par an; no topereshathice
Wil now witheed upper massithis com?

https://nextjournal.com/jbowles/julia-version-of-goldbergs-the-unreasonable-effectiveness-of-character-level-language-models/#generate-shakespeare-from-different-order-models

k-tuple (character) Model of Shakespeare (*k*=9)

First Citizen:
We'll sure speak, and dear men
Of estimations; the
people cry 'Troilus'? Helenus beholds
A Grecian's life did look on.

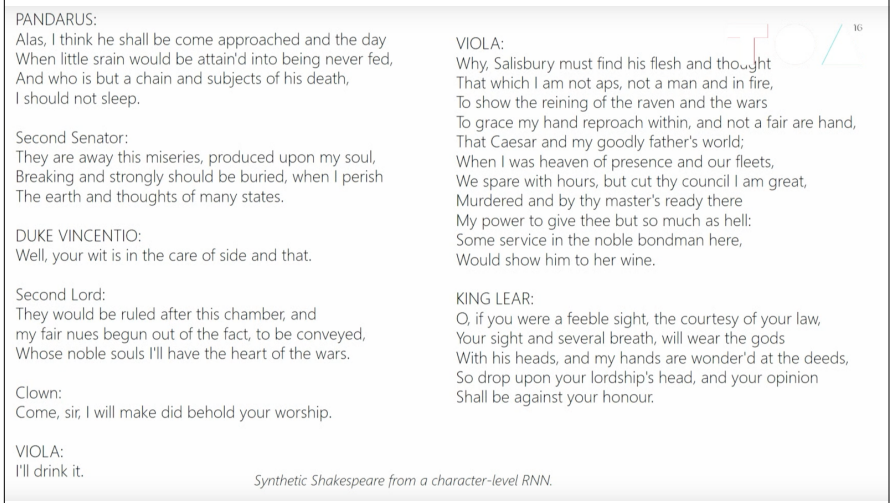
MISTRESS QUICKLY:
And have my eight shillings to one it please your majesty? How fares my Kate? What, are they call'd unto a cardinal,
I should break.

PRINCE HENRY:
Where have you
looked sadly, it was the day!

PROSPERO:
Thou and my brother:--and she hath importuned you;
Nor now my business in thine answer.

https://nextjournal.com/jbowles/julia-version-of-goldbergs-the-unreasonable-effectiveness-of-character-level-language-models/#generate-shakespeare-from-different-order-models

"Facebook's neural network beat the infinite number of monkeys — and wrote Shakespeare"



PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.

VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Synthetic Shakespeare from a character-level RNN.

https://toa.life/how-facebooks-neural-network-beat-the-infinite-number-of-monkeys-and-wrote-shakespeare-a14a1484e7

Using Character and Word-Level Models Improves Quality of "Shakespeare"

Thy youth's time and face his form shall cover?
Now all fresh beauty, my love there
Will ever Time to greet, forget each, like ever decease, But in a best at worship his glory die.

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762063.pdf

Shakespeare is Not Really Modern English

"My life itself, and the best heart of it,
Thanks you for this great care: I stood i' the level
Of a full-charg'd confederacy, and give thanks
To you that chok'd it. Let be call'd before us
That gentleman of Buckingham's; in person
I'll hear him his confessions justify;
And point by point the treasons of his master
He shall again relate." —*Henry VIII*

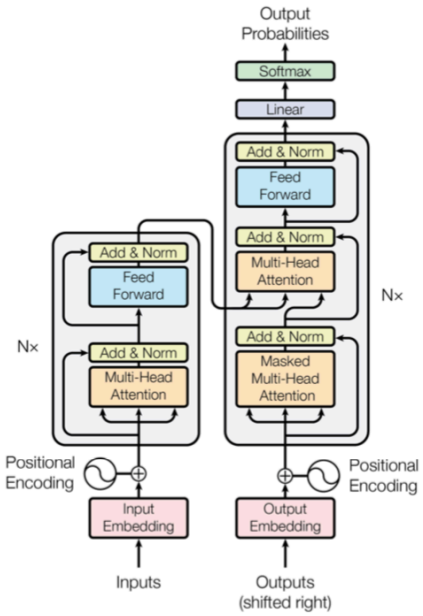


Figure 1: The Transformer - model architecture.

"Attention is All You Need" (Transformer)

- Seq-Seq basic idea
- "Self-attention"
- Residual connections
- Normalization at each layer
- Multi-head attention (like multi-kernel convolution)
  - Only attend to outputs already generated
- Position encoding

http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

BERT
Bidirectional Encoder Representations from Transformers

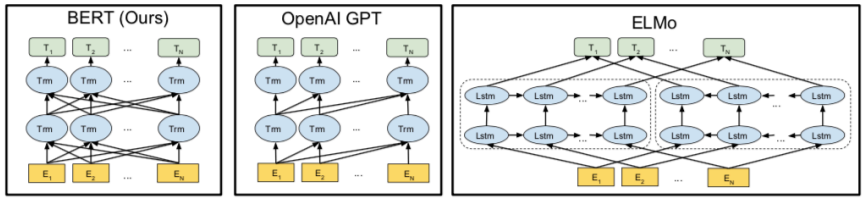


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- Word-piece tokens
- Predict masked tokens (~15%)
- Predict next sentence
- Trained on 800M word Books, 2,500M word Wikipedia corpus
- Large performance improvement on many tasks

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 10). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

OpenAI's GPT-2 Model
(Generative Pre-trained Transformer)

- Transformer-based
- Trained on 8M Web pages
- Model predicts next word (simple)
- Consistently outperforms systems trained on more specific corpora, e.g., Wikipedia, news, books
- "Fake news" generator

https://openai.com/blog/better-language-models

GPT-2 Example
(1st try for prompt)

Prompt

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.

"The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses.

https://openai.com/blog/better-language-models

GPT-2 Example (10th try for prompt)

Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the

https://openai.com/blog/better-language-models

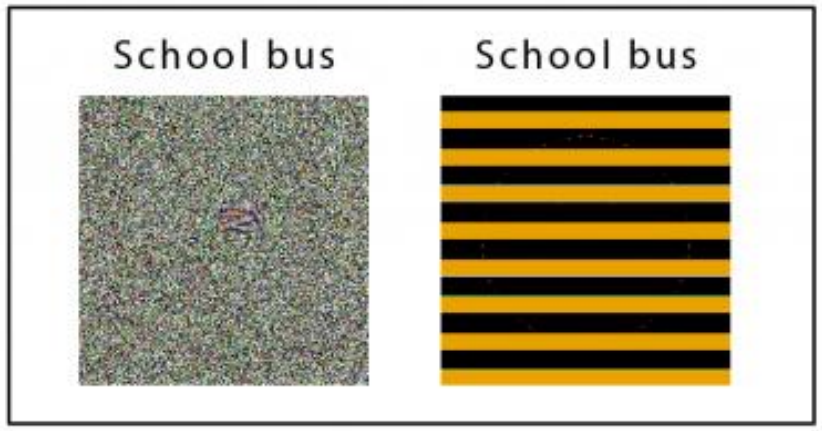
How Good is this Model?

- Language use:
  - Amazing
- Logic:
  - "the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans"

Before We Get Carried Away…

- There is no "gestalt"
  - "form" or "shape"
  - Holism: "natural systems and their properties must be viewed as wholes, not as loose collections of parts"
  - Instead, attention to many features

Szegedy et al. 2014

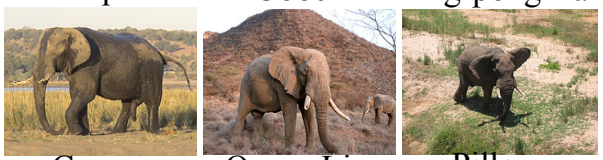



Ocean Liner Ocean Liner Ocean Liner

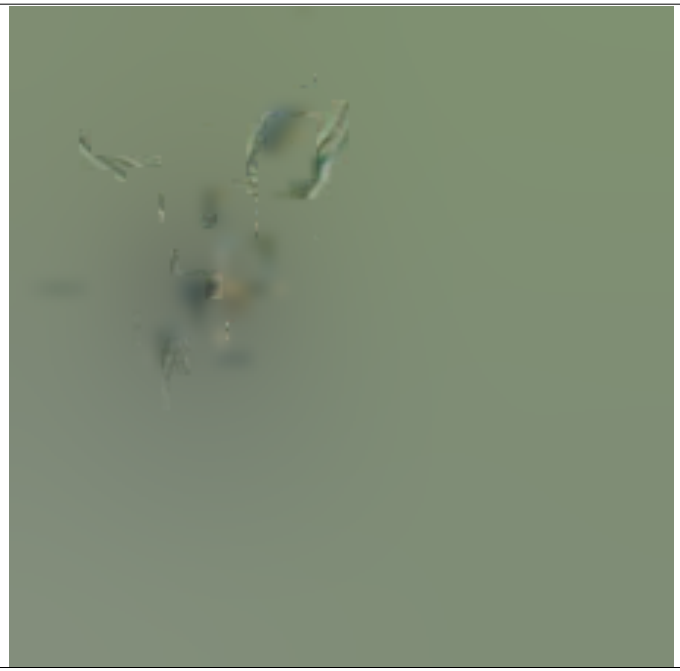

Elephant Elephant Elephant


House House House

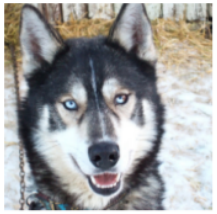

Cup Oboe Ping-pong Ball


Canon Ocean Liner Pillow


Envelope Centipede Fly



Husky vs. Wolf



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf