EN. 601.647/667 INTRODUCTION TO HLT DEEP LEARNING

KENTON MURRAY

9/22/2026

EN. 601.647/667 INTRODUCTION TO HLT DEEP LEARNING

KENTON MURRAY

9/22/2026

Some slides were inspired by:
Kevin Duh, Shinji Watanabe, Marine Carpaut, Graham Neubig, Jacob Eisenstein

I'M HIRING PHD STUDENTS

- https://docs.google.com/forms/d/e/1FAIpQLScZqqr7xN63lpP7m7Z22_oEVa257f7_jJp4YOBTDURfWck1JA/viewform?usp=sharing&ouid=117382557743098706243

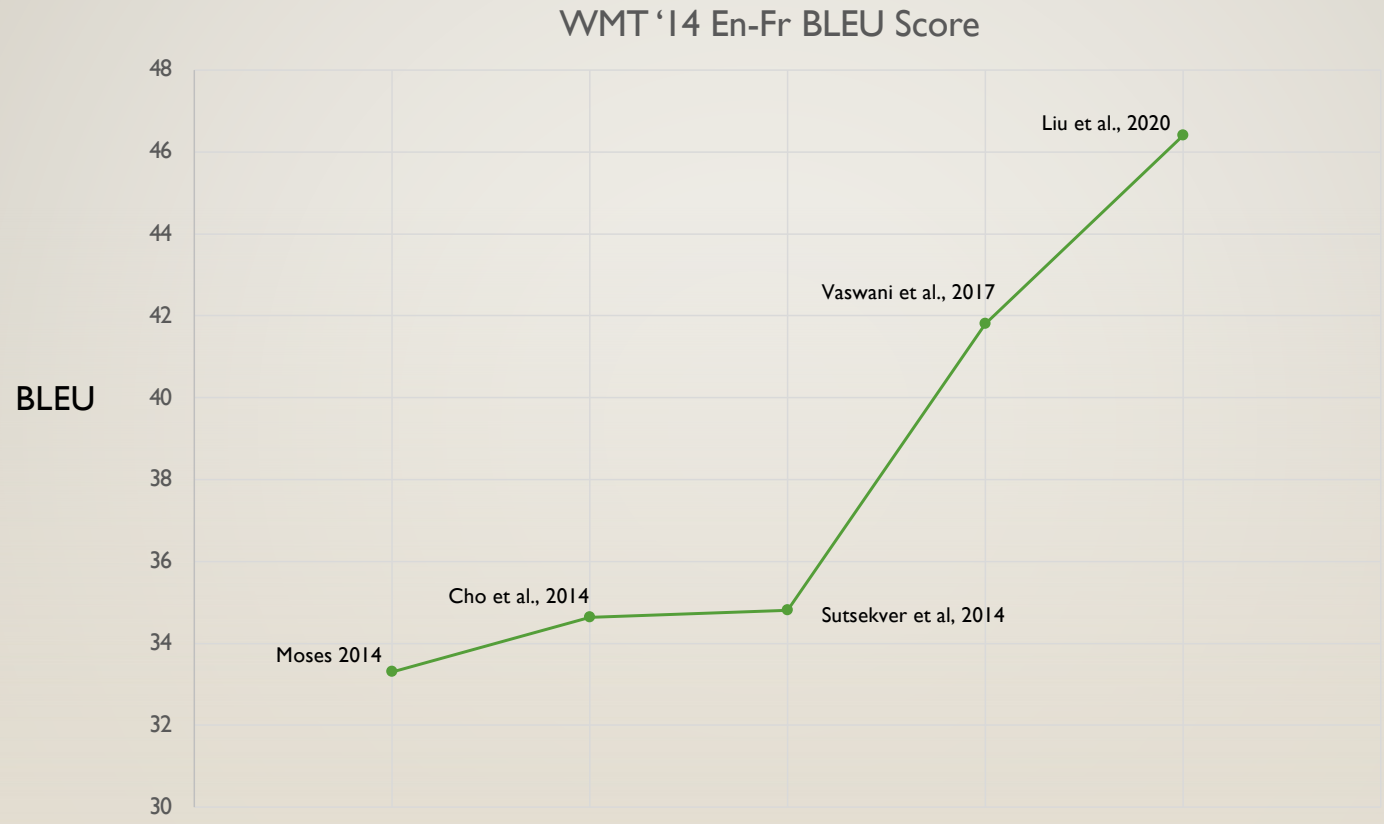
AGENDA

- Impact of Deep Learning on HLT
- Intro to Deep Learning
- What changed? Why recently?
- Deep Learning in HLT
- PyTorch

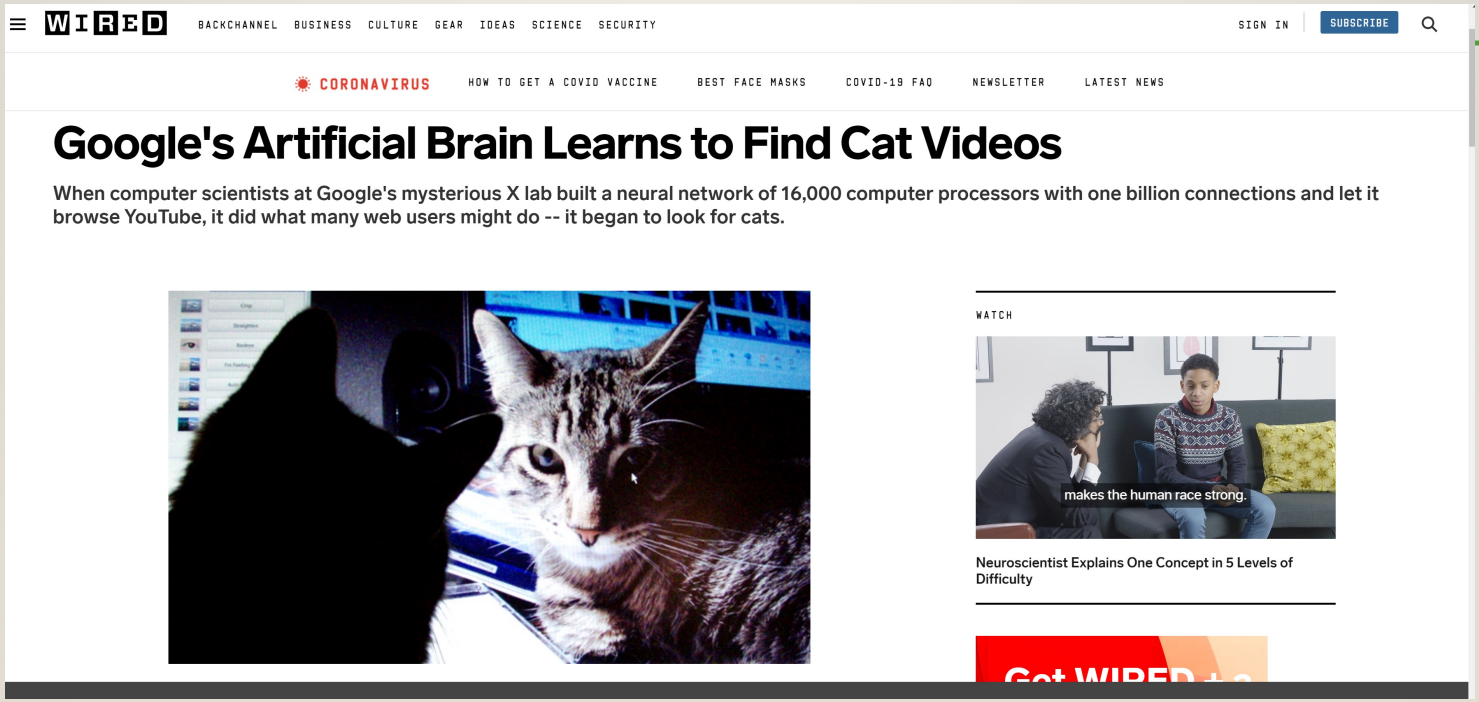
QUICK ASIDE

- BLEU Scores
- Modified n-gram precision metric for Machine Translation
- 0 – 100.0 (higher is better)
- Reviewers generally like ~+1.0 gain over a baseline

IMPACT OF DEEP LEARNING



CATS!


WIRED
BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY
SIGN IN
SUBSCRIBE
CORONAVIRUS
HOW TO GET A COVID VACCINE
BEST FACE MASKS
COVID-19 FAQ
NEWSLETTER
LATEST NEWS
Google's Artificial Brain Learns to Find Cat Videos
When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.
WATCH
makes the human race strong.
Neuroscientist Explains One Concept in 5 Levels of Difficulty

ACL 2014 BEST PAPER

Fast and Robust Neural Network Joint Models for Statistical Machine Translation

Jacob Devlin, Rabih Zbib, Zhongqiang Huang,
Thomas Lamar, Richard Schwartz, and **John Makhoul**
Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA 02138, USA
{jdevlin,rzbib,zhuang,tlamar,schwartz,makhoul}@bbn.com

Abstract

Recent work has shown success in using neural network language models (NNLMs) as features in MT systems. Here, we present a novel formulation for a neural network *joint* model (NNJM), which augments the NNLM with a source context window. Our model is purely lexicalized and can be integrated into any MT decoder. We also present several variations of the NNJM which provide significant additive improvements.

Although the model is quite simple, it

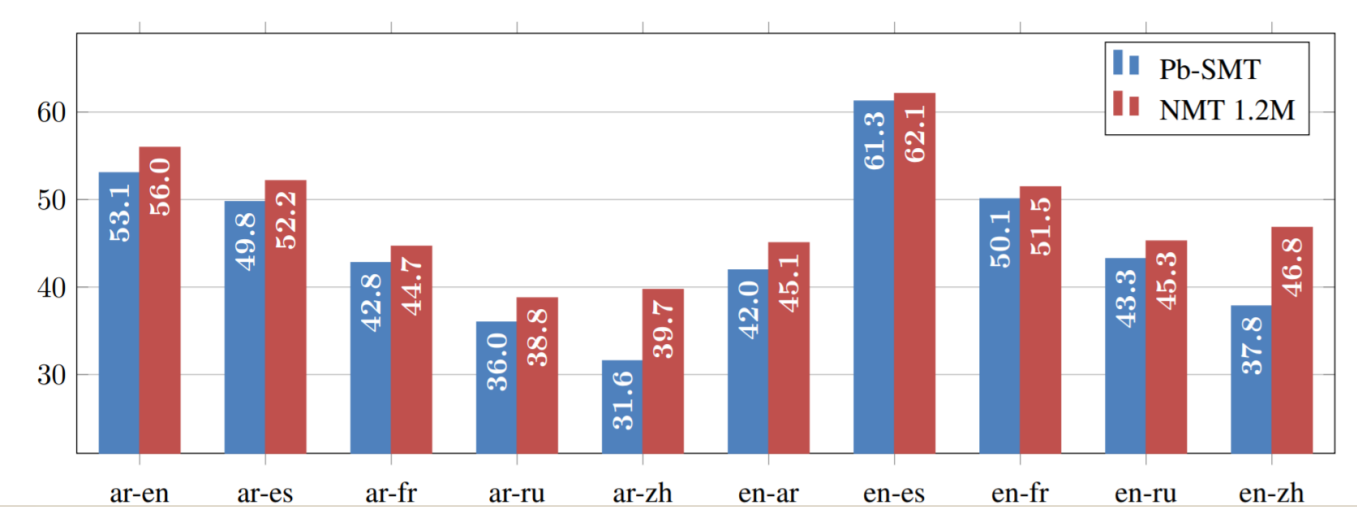
1 Introduction

In recent years, neural network models have become increasingly popular in NLP. Initially, these models were primarily used to create n-gram neural network language models (NNLMs) for speech recognition and machine translation (Bengio et al., 2003; Schwenk, 2010). They have since been extended to translation modeling, parsing, and many other NLP tasks.

In this paper we use a basic neural network architecture and a lexicalized probability model to create a powerful MT decoding feature. Specifically, we introduce a novel formulation for a neu-

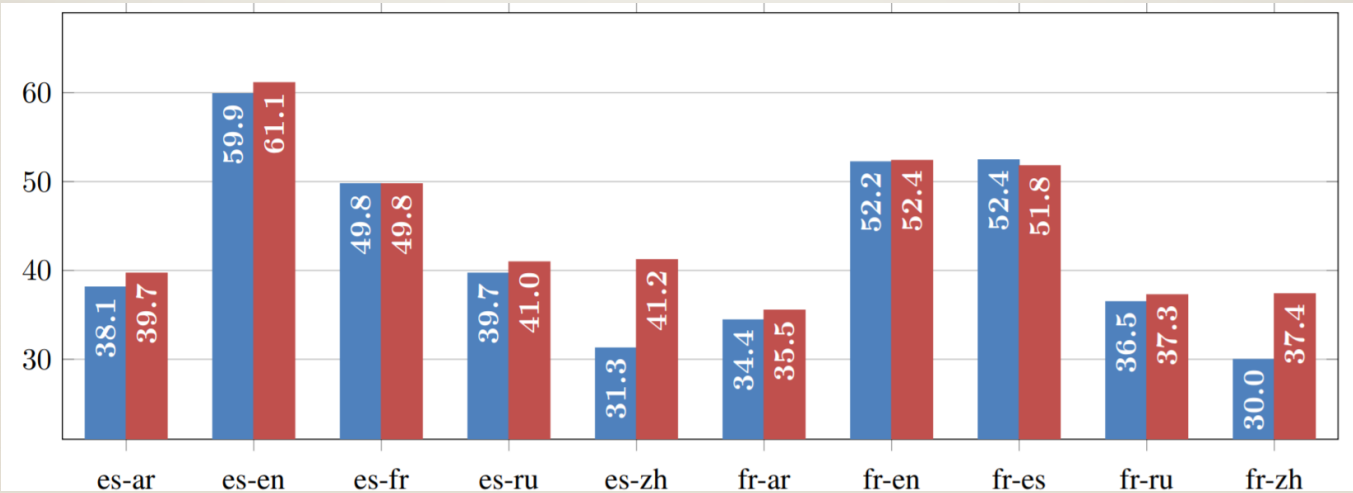
IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

- Junczys-Dowmunt et al., 2016



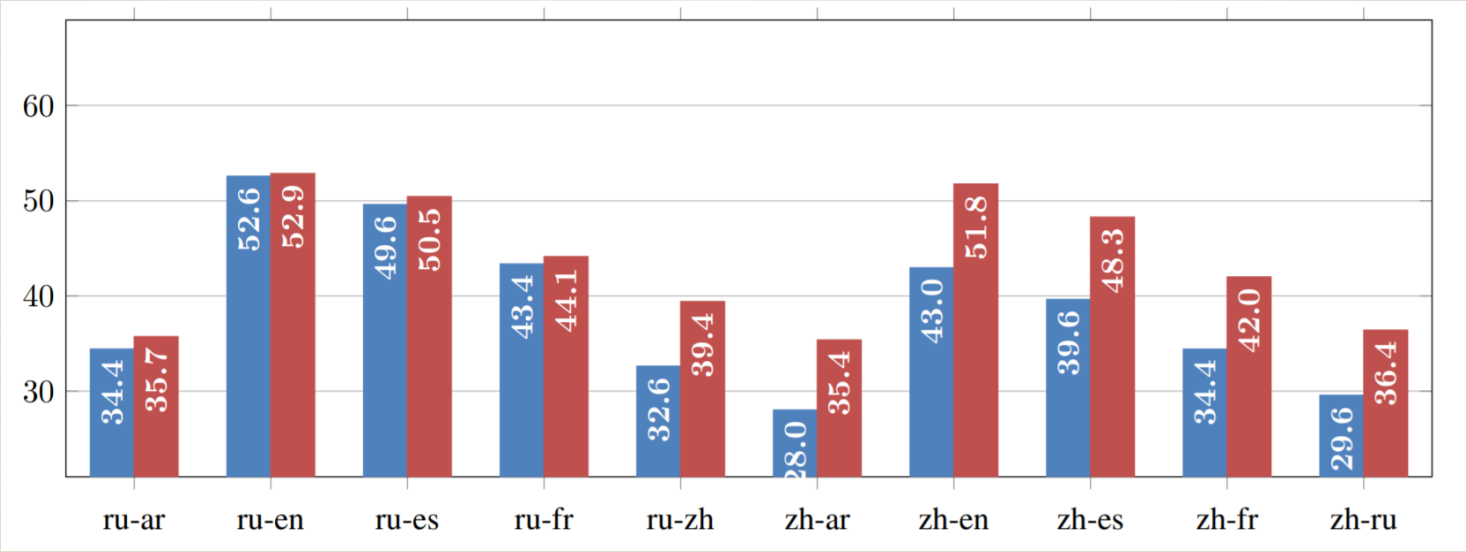
IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

- Junczys-Dowmunt et al., 2016



IS NEURAL MACHINE TRANSLATION READY FOR DEPLOYMENT?

- Junczys-Dowmunt et al., 2016



DEEP LEARNING BACKGROUND

- Deep Learning is Neural Networks ….That are Deep!

SUPERVISED LEARNING

- Model: θ
- Input: X
- Output: Y
- $P_\theta(Y|X)$

SUPERVISED LEARNING (LANGUAGE ID)

- Model: θ
- Input: X (Hello my name is)
- Output: Y (English)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (LANGUAGE ID)

- Model: θ
- Input: X (Hola, mi nombre es)
- Output: Y (Spanish)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (LANGUAGE ID)

- Model: θ
- Input: X (Salut, je m'appelle)
- Output: Y (French)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (LANGUAGE ID)

- Model: θ
- Input: X (Imanalla, nuqap suti Murray kan)
- Output: Y (?)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (SPEECH RECOGNITION)

- Model: θ
- Input: X ()
- Output: Y (?)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (SPEECH RECOGNITION)

- Model: θ
- Input: X ()
- Output: Y (Hello World)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (MACHINE TRANSLATION)

- Model: θ
- Input: X (مرحبا بالعالم)
- Output: Y (?)
- $P_\theta(Y|X)$

SUPERVISED LEARNING (MACHINE TRANSLATION)

- Model: θ
- Input: X (مرحبا بالعالم)
- Output: Y (Hello World)
- $P_\theta(Y|X)$

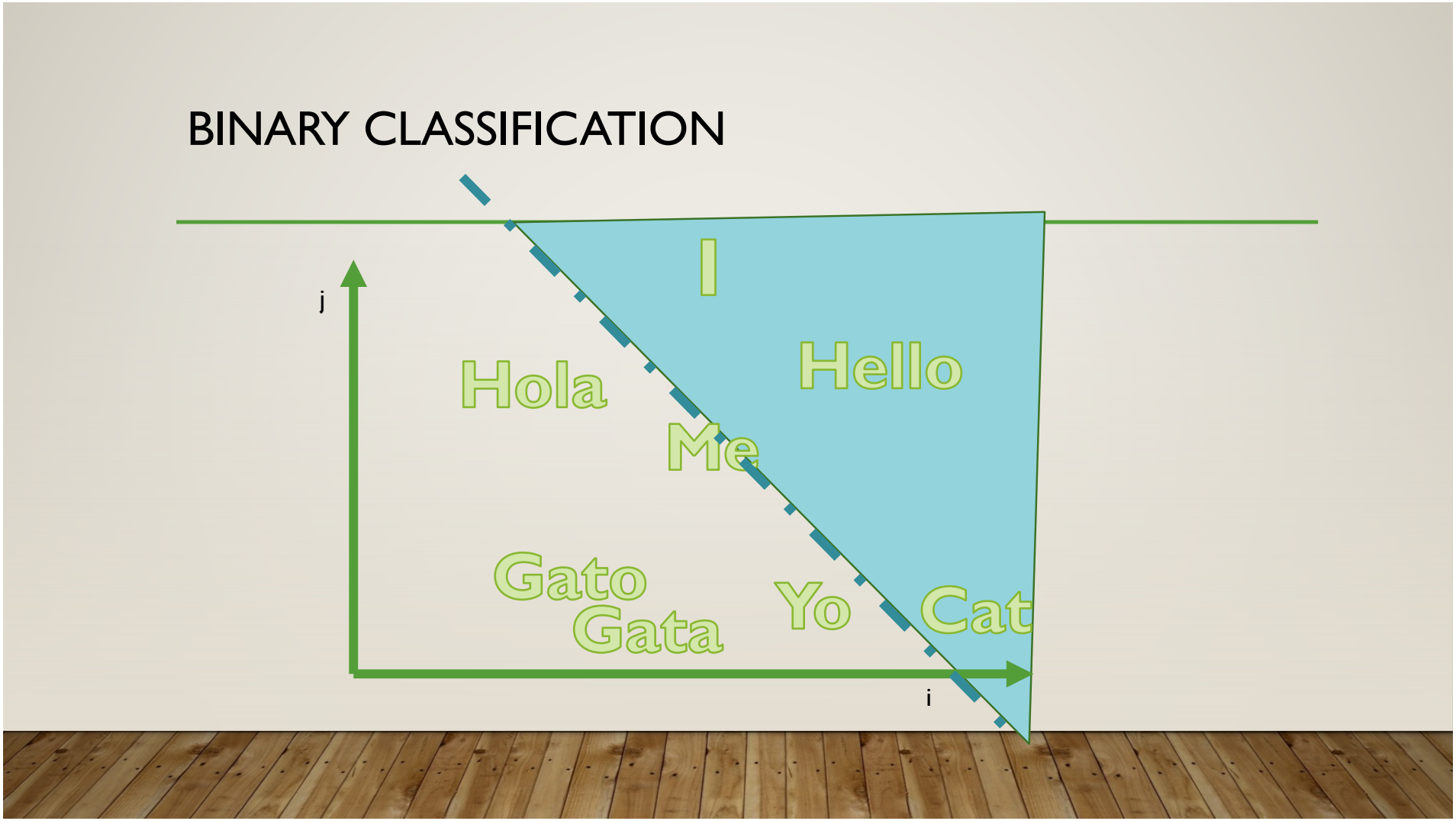
CLASSIFICATION

- One of the first tasks public thinks of
- Sentiment Analysis
- Speaker/Writer ID
- Language ID
- Phoneme Recognition

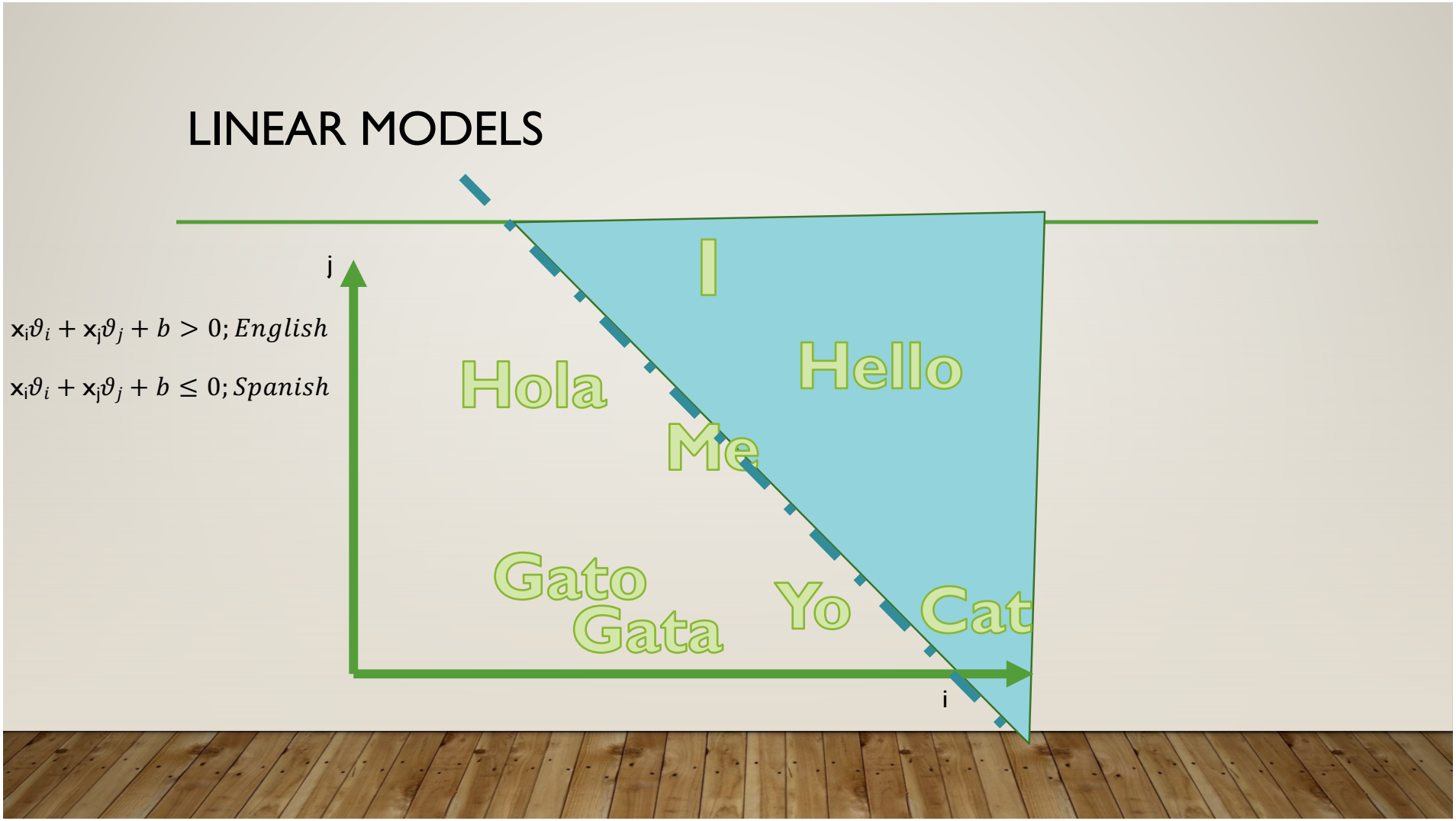
BINARY CLASSIFICATION

BINARY CLASSIFICATION

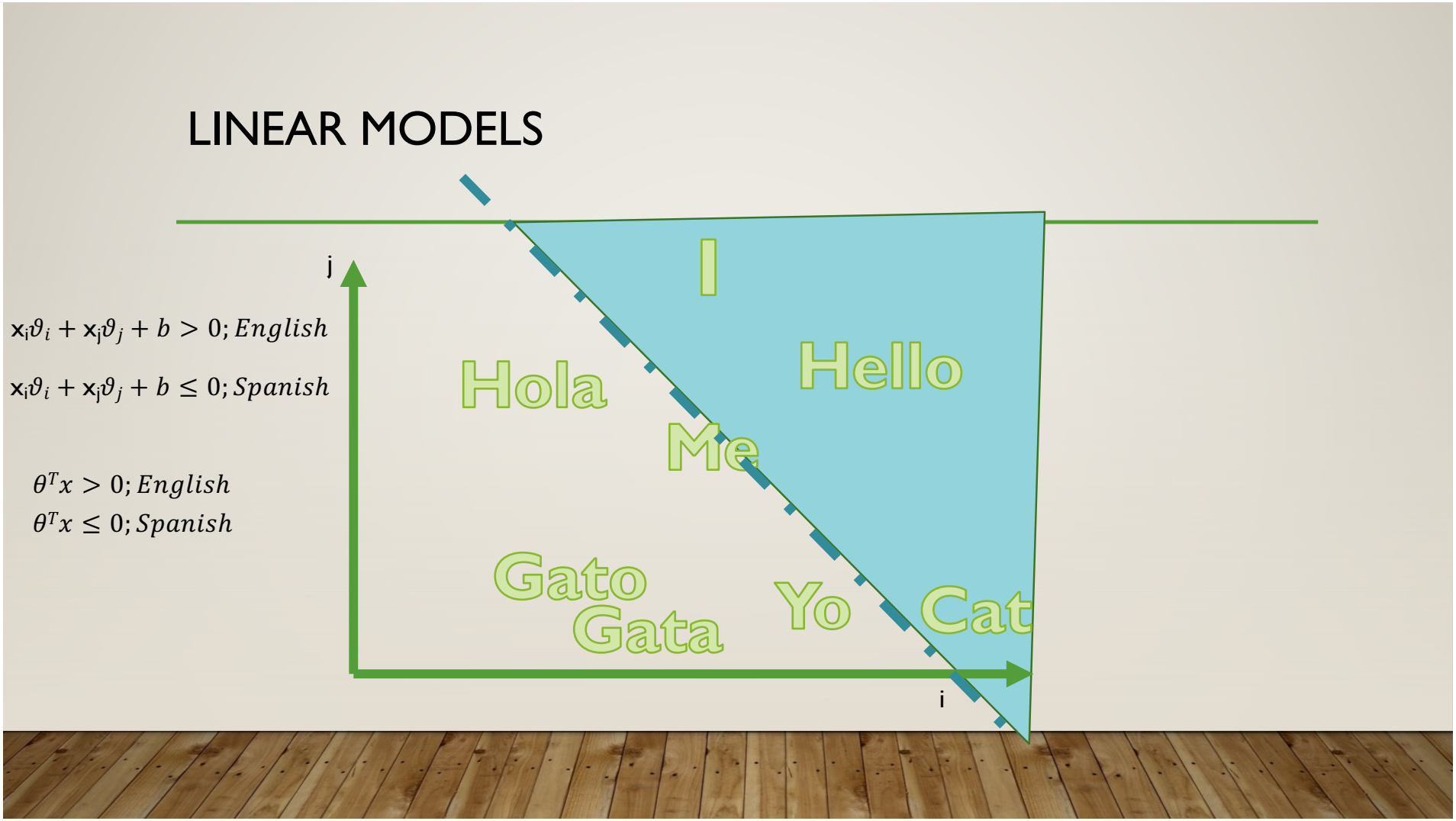
BINARY CLASSIFICATION



LINEAR MODELS

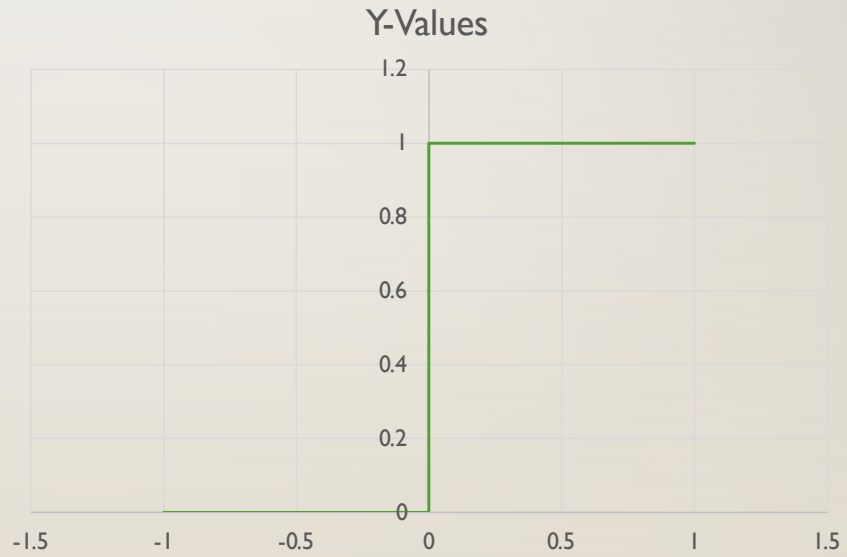


LINEAR MODELS



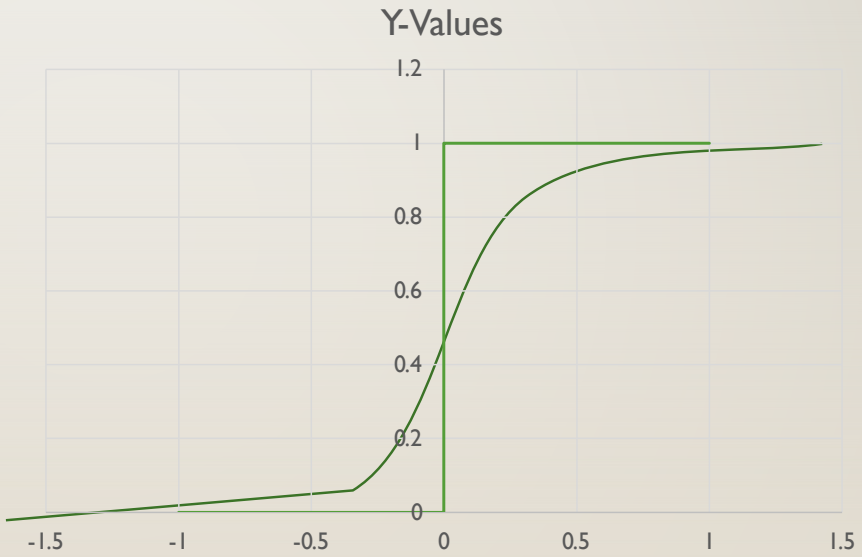
PERCEPTRON

- Make it a probability
- P(y=English|x) = 1.0 if $\theta^T x > 0$
- P(y=English|x) = 0.0 if $\theta^T x \leq 0$



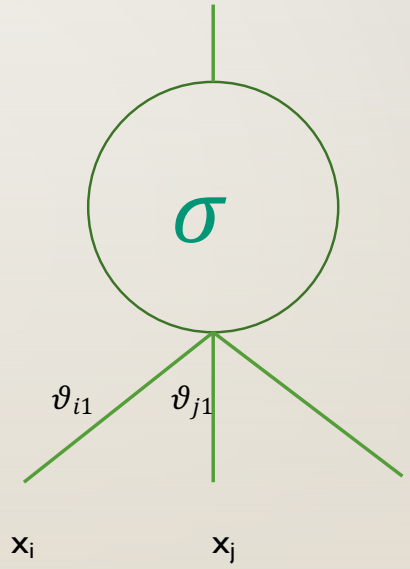
LOGISTIC REGRESSION

- Make it a probability
- P(y=English|x) = $\sigma\ (\theta^T x)$
- P(y=English|x) = $\frac{1}{1+e^{\theta Tx}}$
- Softer
- Differentiable

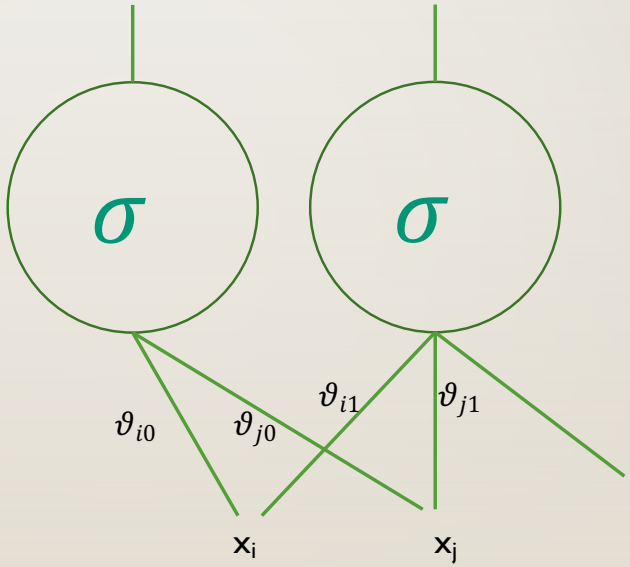


LOGISTIC REGRESSION

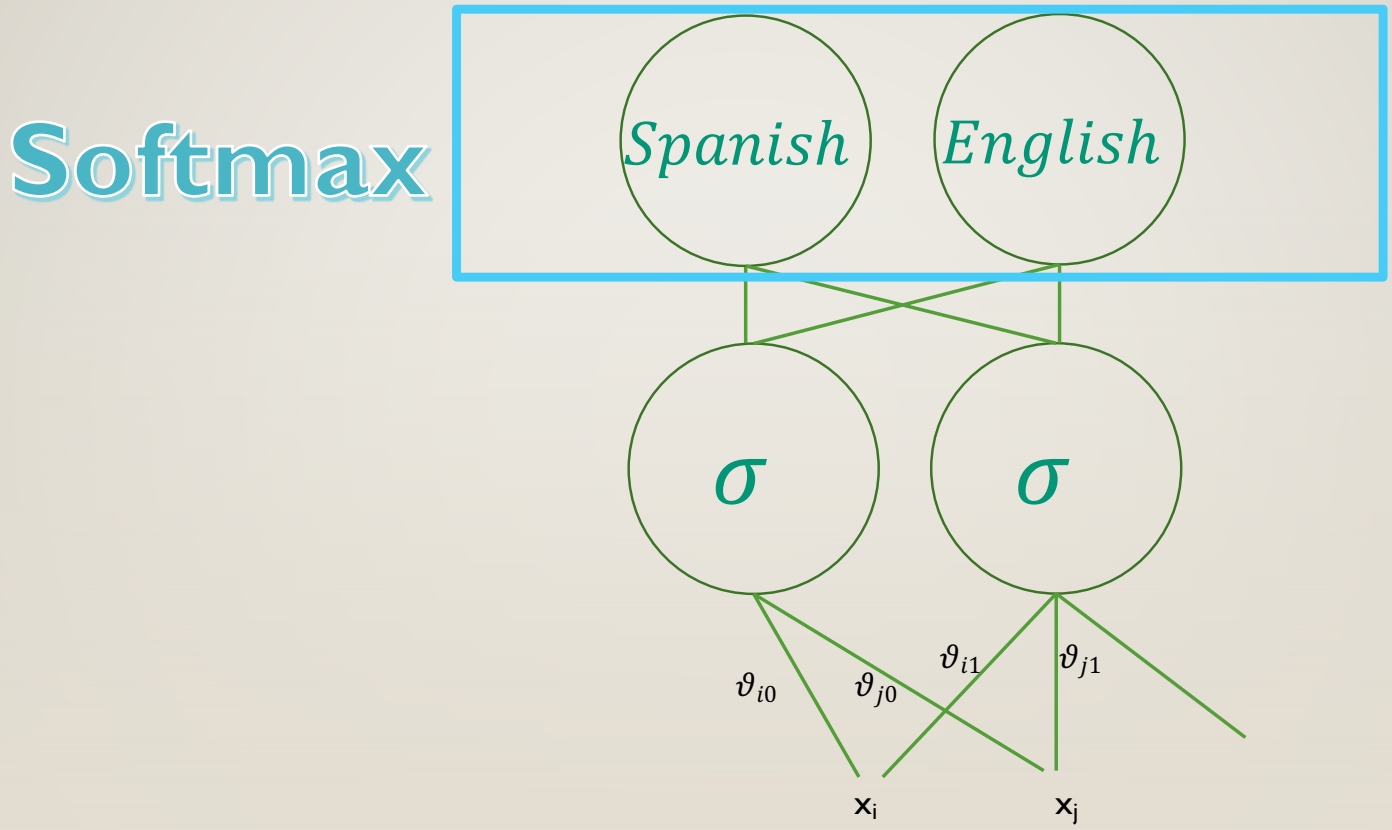
- Make it a probability
- P(y=English|x) = $\sigma\ (\theta^T x)$
- P(y=English|x) = $\frac{1}{1+e^{\theta T x}}$
- Softer
- Differentiable



2 NEURON FEED-FORWARD



2 Neuron Feed-Forward



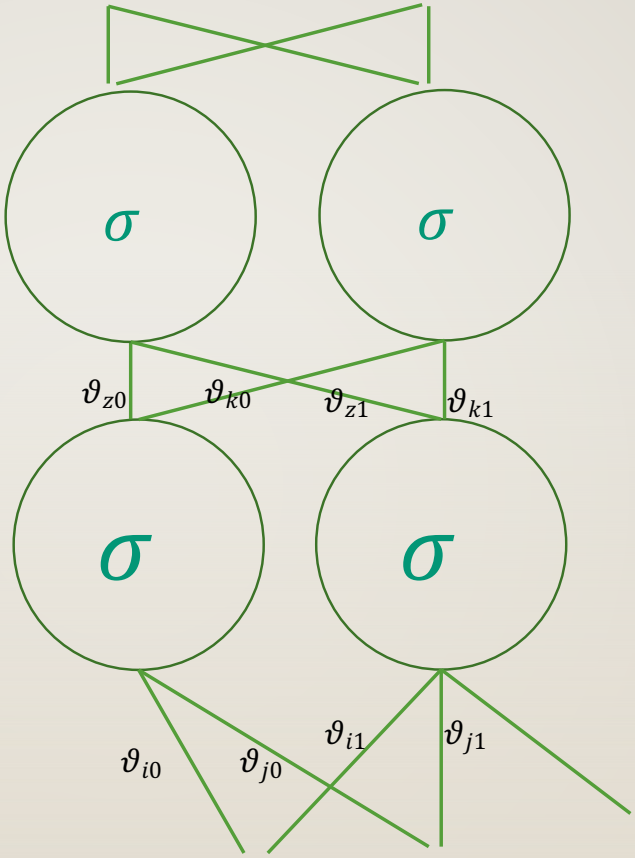
SOFTMAX

- Make the output a probability distribution
- $\sigma\,(z_i) = \frac{e^{zi}}{\sum_{j=0}^{Z} e^{z_j}}$

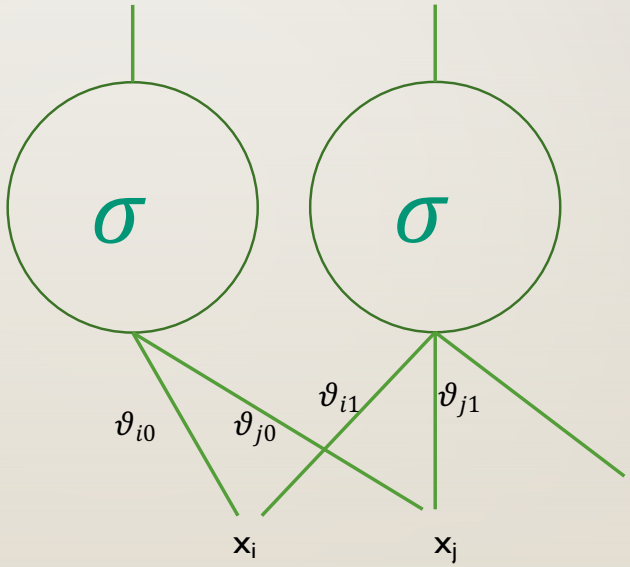
SOFTMAX

- Make the output a probability distribution
- P(y=English|x) = $\frac{1}{1+e^{\theta Tx}}$
- $\sigma\ (z_i) = \frac{e^{zi}}{\sum_{j=0}^{Z} e^{z_j}}$
- Training: Differentiable through it
- Testing: Take the Max

DEEP

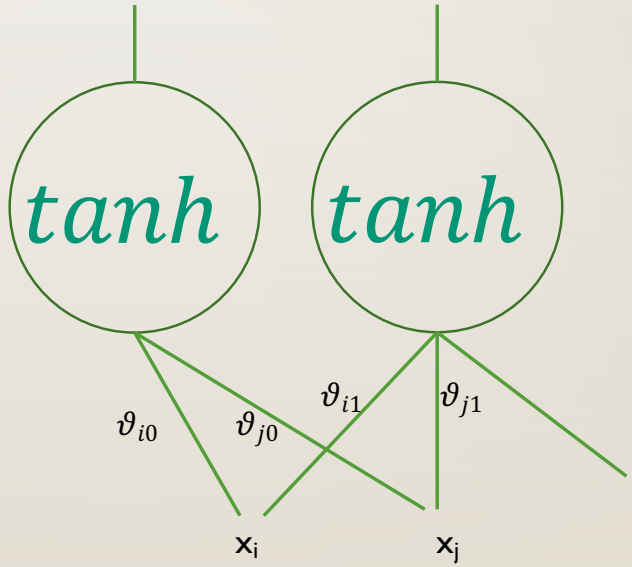

σ
σ
ϑ_{z0}
ϑ_{k0}
ϑ_{z1}
ϑ_{k1}
σ
σ
ϑ_{i0}
ϑ_{j0}
ϑ_{i1}
ϑ_{j1}

OTHER NON-LINEARITIES



OTHER NON-LINEARITIES

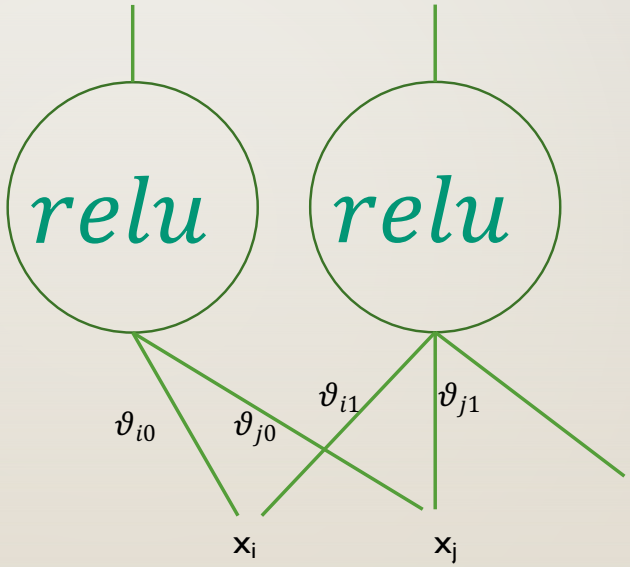
- A lot like sigmoid
- Range: (-1.0, 1.0)



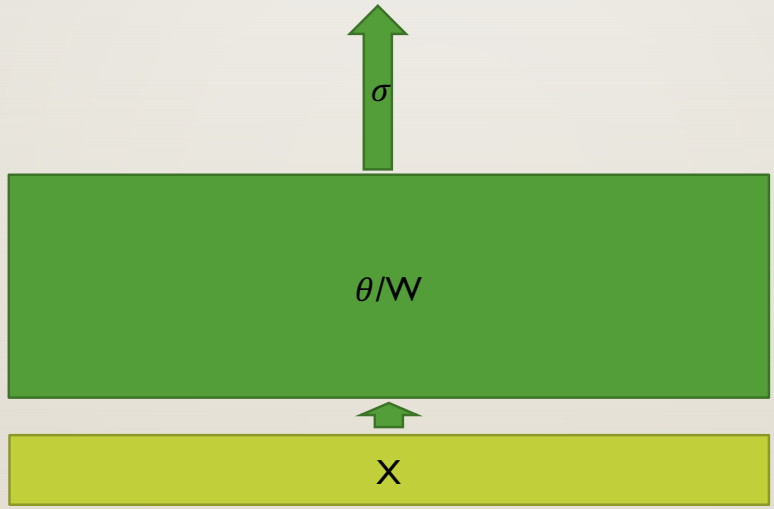
OTHER NON-LINEARITIES

relu(x)= $0; if\ x < 0$

relu(x)= $x; if\ x \geq 0$



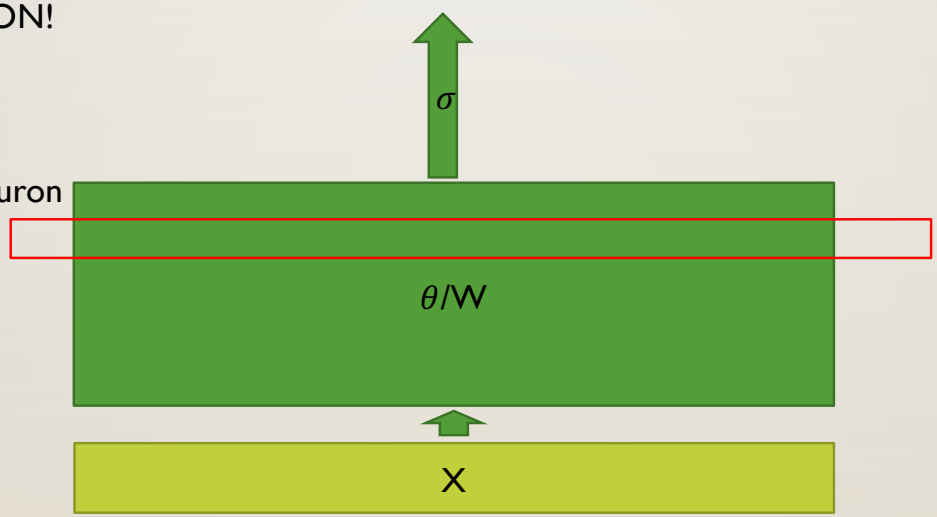
TENSORS!



TENSORS!

DEPENDS ON DEFINITION!

Row: weights into a single neuron

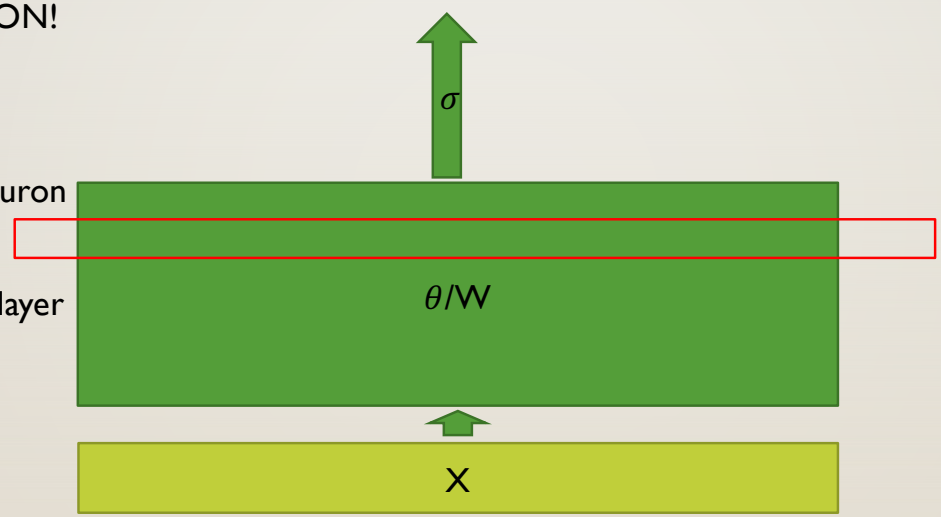


TENSORS!

DEPENDS ON DEFINITION!

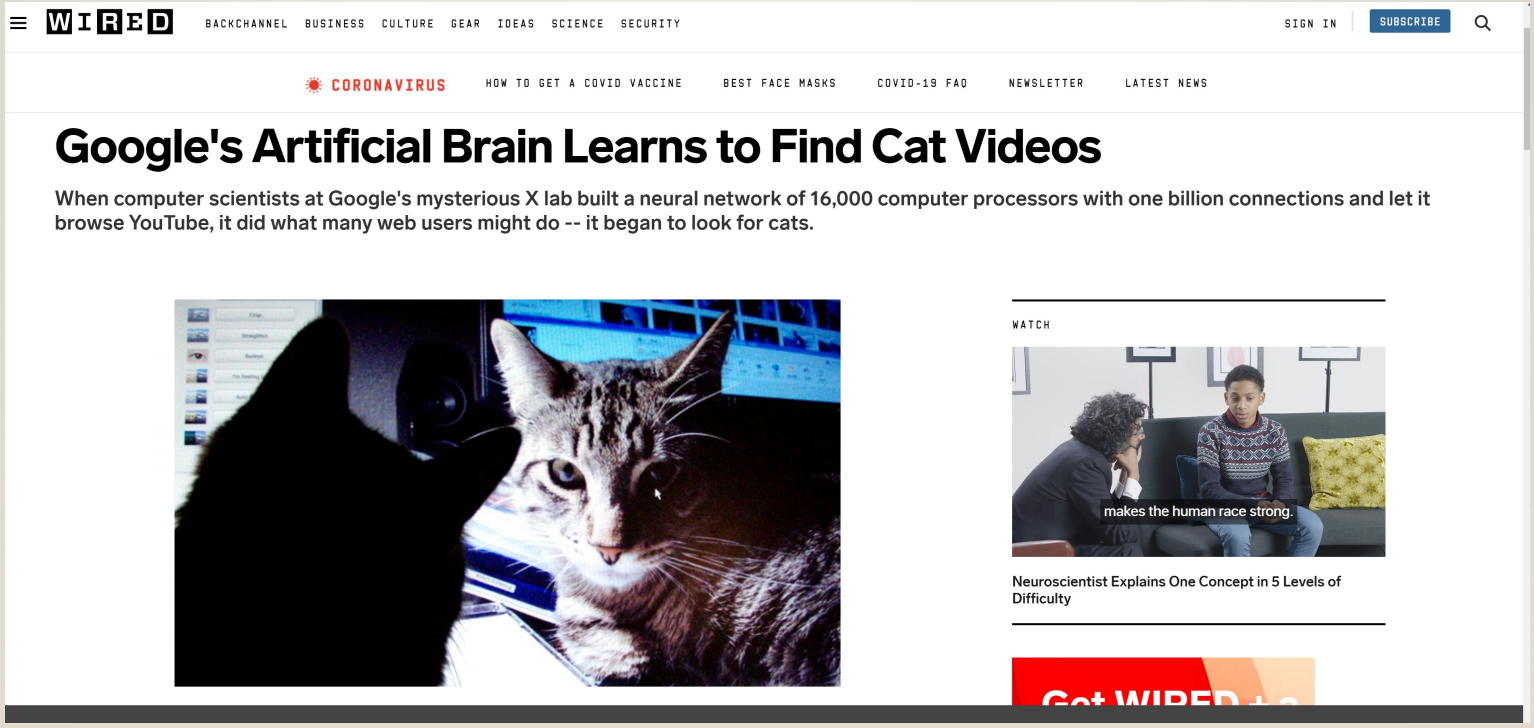
Row: weights into a single neuron

Cols: Number of neurons in layer

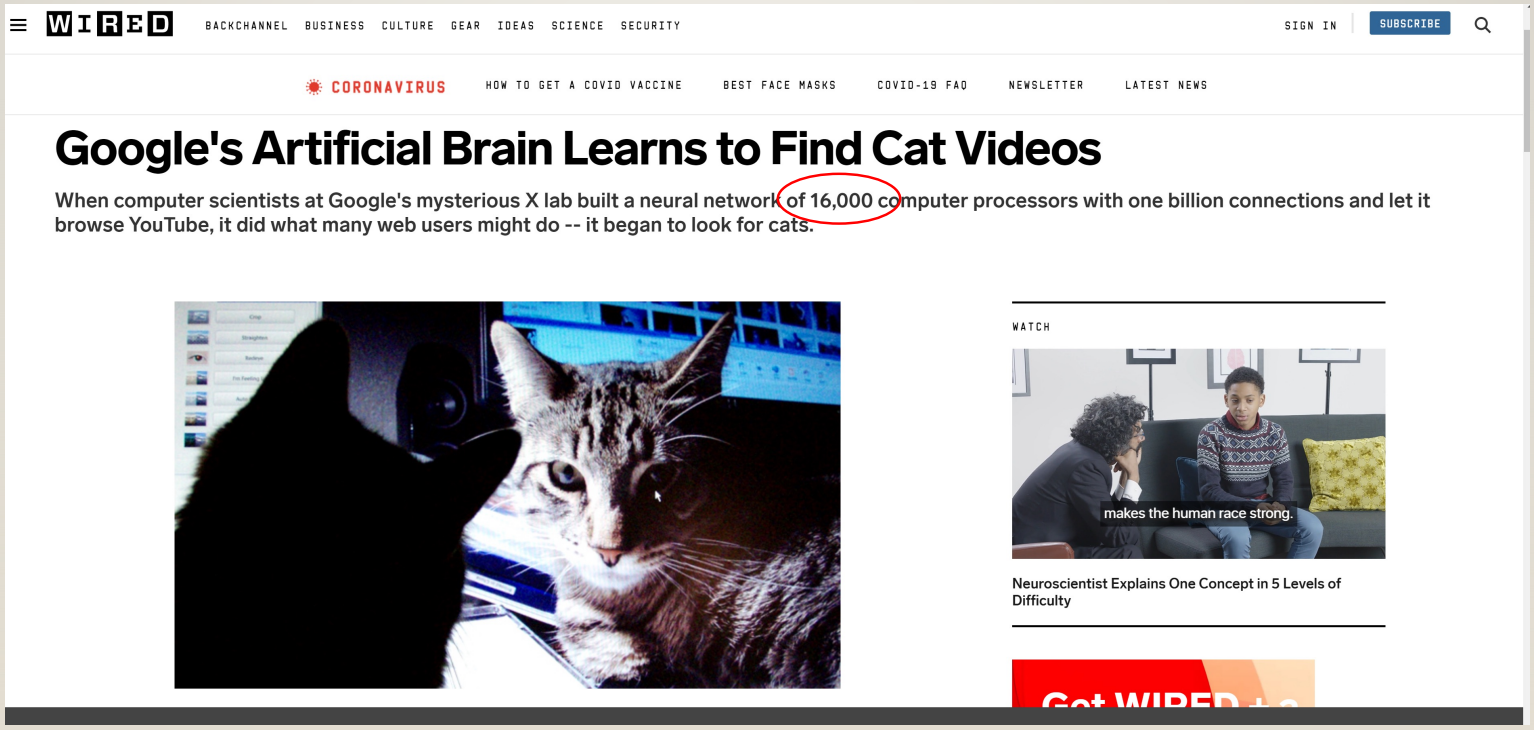


WHY NOW? (LAST 10 YEARS)

CATS!


WIRED
BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY
SIGN IN
SUBSCRIBE
CORONAVIRUS
HOW TO GET A COVID VACCINE
BEST FACE MASKS
COVID-19 FAQ
NEWSLETTER
LATEST NEWS
Google's Artificial Brain Learns to Find Cat Videos
When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.
WATCH
makes the human race strong.
Neuroscientist Explains One Concept in 5 Levels of Difficulty

CATS!


WIRED
BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY
SIGN IN
SUBSCRIBE
CORONAVIRUS
HOW TO GET A COVID VACCINE
BEST FACE MASKS
COVID-19 FAQ
NEWSLETTER
LATEST NEWS
Google's Artificial Brain Learns to Find Cat Videos
When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.
WATCH
makes the human race strong.
Neuroscientist Explains One Concept in 5 Levels of Difficulty

CATS!

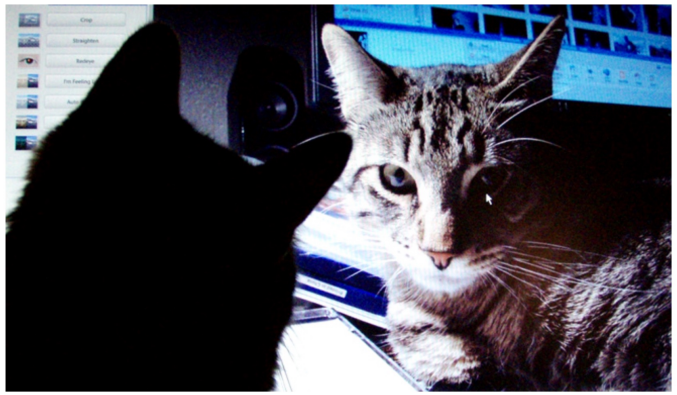
GPUs!

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

CORONAVIRUS HOW TO GET A COVID VACCINE BEST FACE MASKS COVID-19 FAQ NEWSLETTER LATEST NEWS

Google's Artificial Brain Learns to Find Cat Videos

When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do -- it began to look for cats.



WATCH

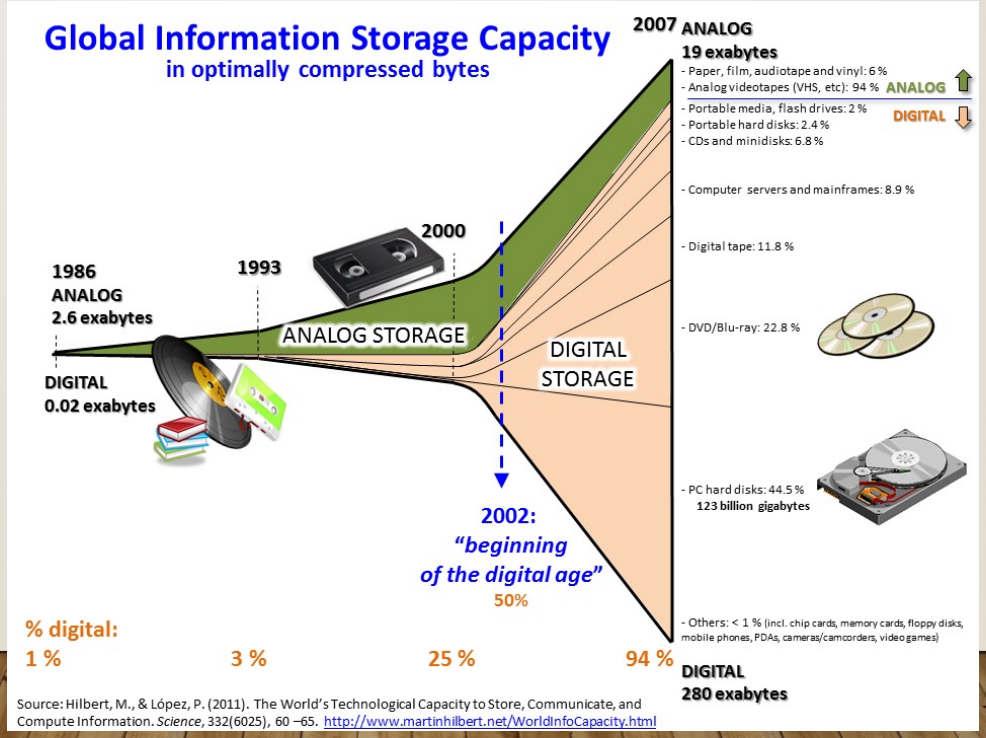
Neuroscientist Explains One Concept in 5 Levels of Difficulty

WHY NOW? (LAST 10 YEARS)

- Murray and Chiang, 2015
- 5-gram Language Model ~28 Days CPU (Grid restrictions)
- Today ~Couple of hours GPU

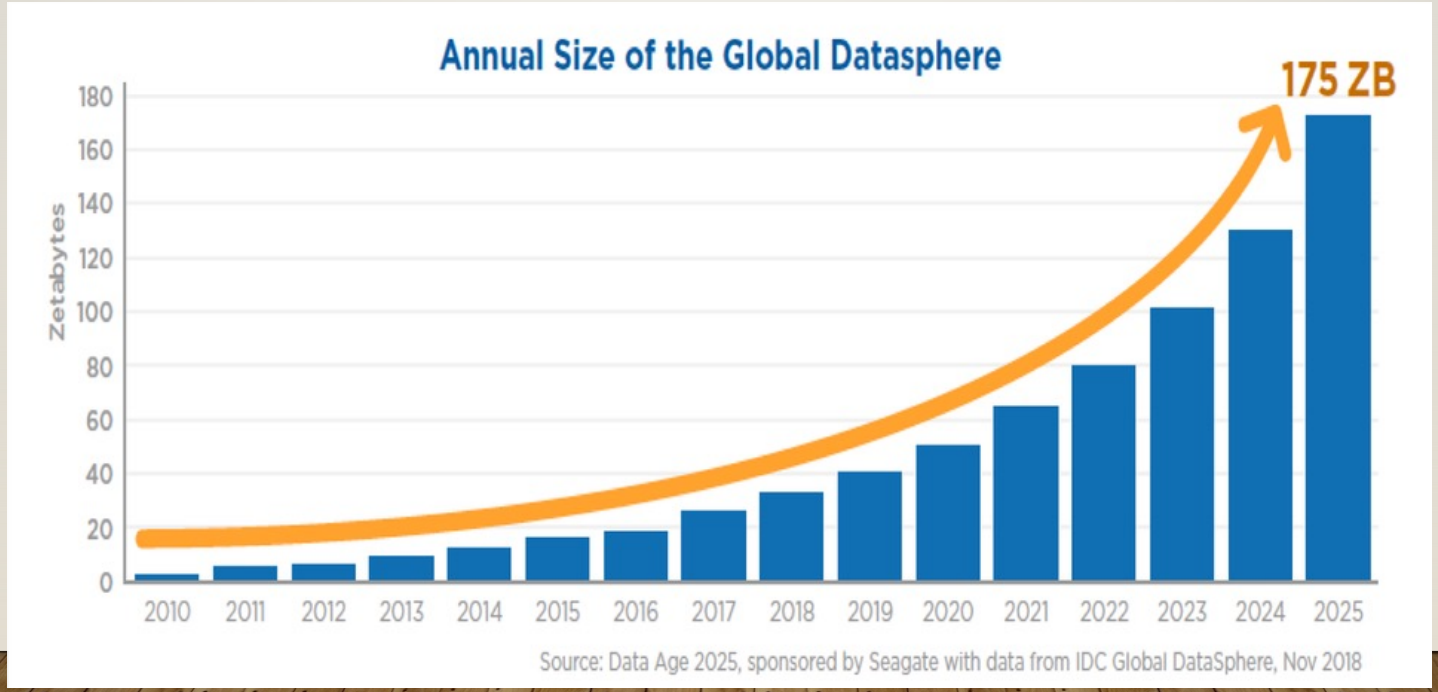
WHY NOW? (LAST 15 YEARS)

- Data availability



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

WHY NOW? (LAST 15YEARS)



Forbes, 2018

LANGUAGE MODELS

- Throwback to first week
- Predict the next word in a sequence

LANGUAGE MODELS

- Throwback to first week
- Predict the next word in a sequence

LANGUAGE MODELS

- Throwback to first week
- Predict the next word in a sequence

Johns Hopkins University was _____

LANGUAGE MODELS

- Historically: vocab, V
- n-gram language model was $|V|^n$
- Data Sparsity issues (5-gram was common)

Johns Hopkins University was _____

MACHINE TRANSLATION

- Source Sentence, f: "Yo tengo hambre"
- Target Sentence, e: "I am hungry"
- P(e|f) …….. Neural Network

MACHINE TRANSLATION

- Cho et al., 2014

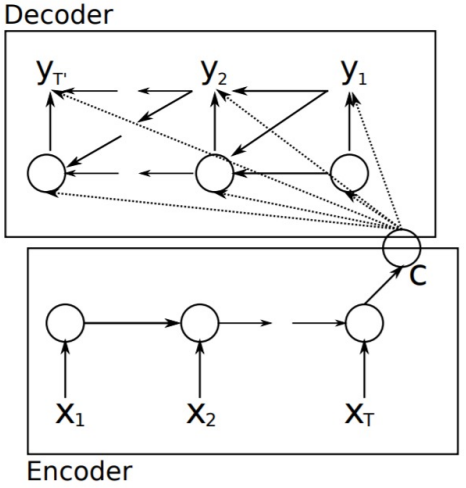


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

MACHINE TRANSLATION

- Cho et al., 2014
- WMT '14 En-Fr

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64

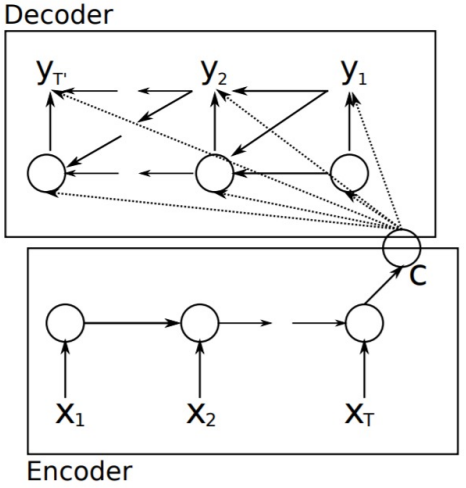


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

MACHINE TRANSLATION

- Sutskever et al., 2014

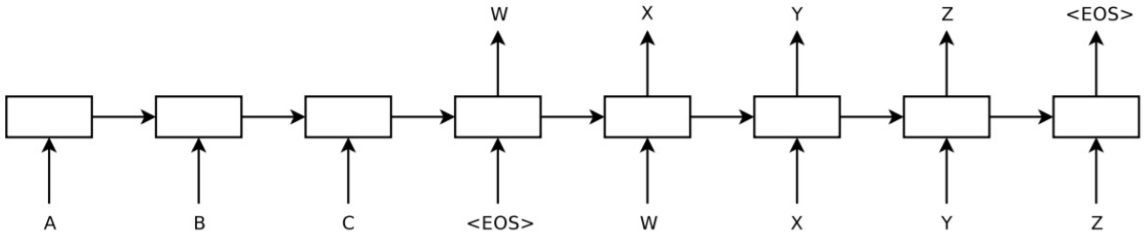
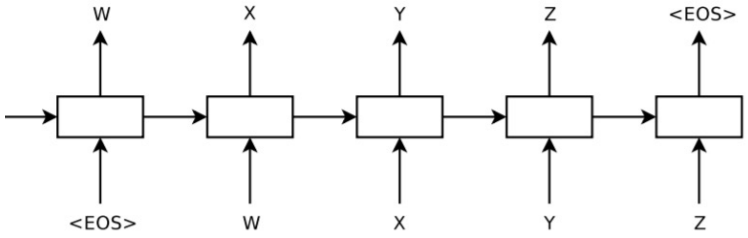


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

MACHINE TRANSLATION

- Sutskever et al., 2014
- WMT '14 En-Fr

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**



entence "ABC" and produces "WXYZ" as the output sentence. The utputting the end-of-sentence token. Note that the LSTM reads the so introduces many short term dependencies in the data that make the

ATTENTION

- Bahdanau et al., 2015

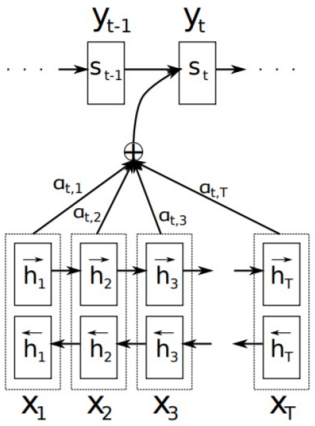


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

ATTENTION

- Bahdanau et al., 2015
- WMT '14 En-Fr

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

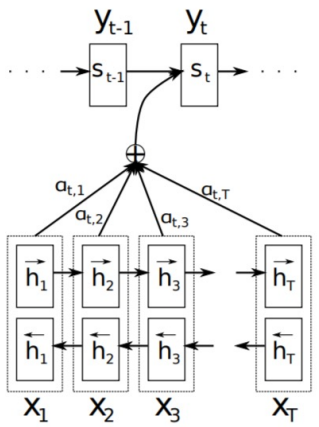
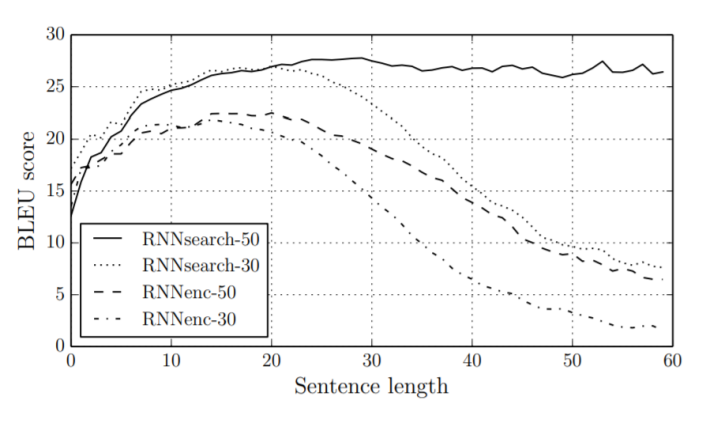


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

ATTENTION

- Bahdanau et al., 2015



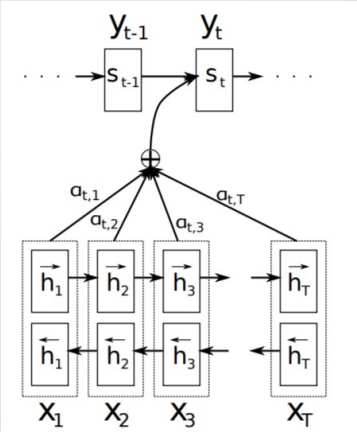


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

ATTENTION IS ALL YOU NEED

- Vaswani et al., 2017
- Transformer

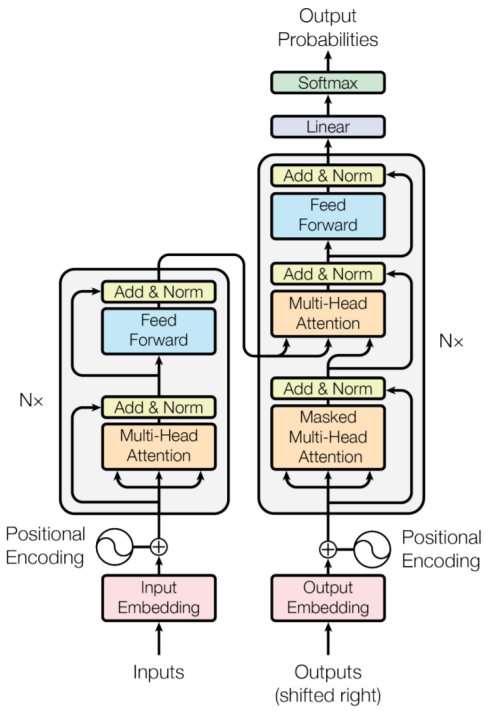


Figure 1: The Transformer - model architecture.

ATTENTION IS ALL YOU NEED

- Vaswani et al., 2017

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

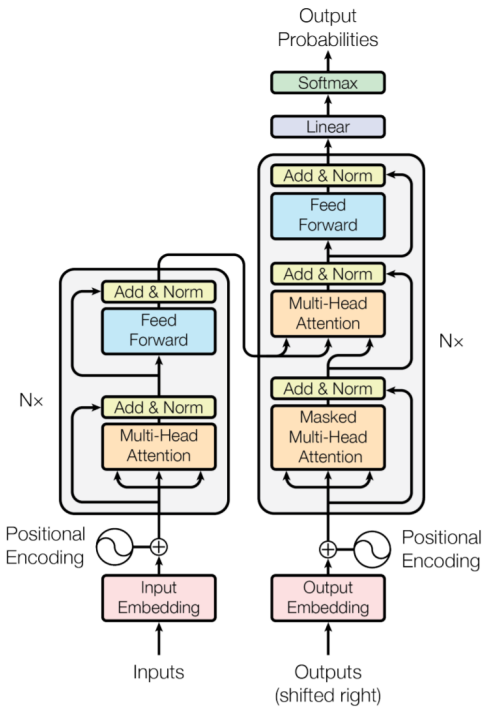


Figure 1: The Transformer - model architecture.

BERT

- Devlin et al., 2018
- Bidirectional Encoder Representations from Transformers
- Masked Language Model

BERT

Johns Hopkins University was _____

BERT

Johns Hopkins University was founded in the year eighteen seventy six .

BERT

Johns Hopkins _____ was founded in the year eighteen _____ six .

Johns Hopkins [ ] was founded in the year eighteen [ ] six .

BERT

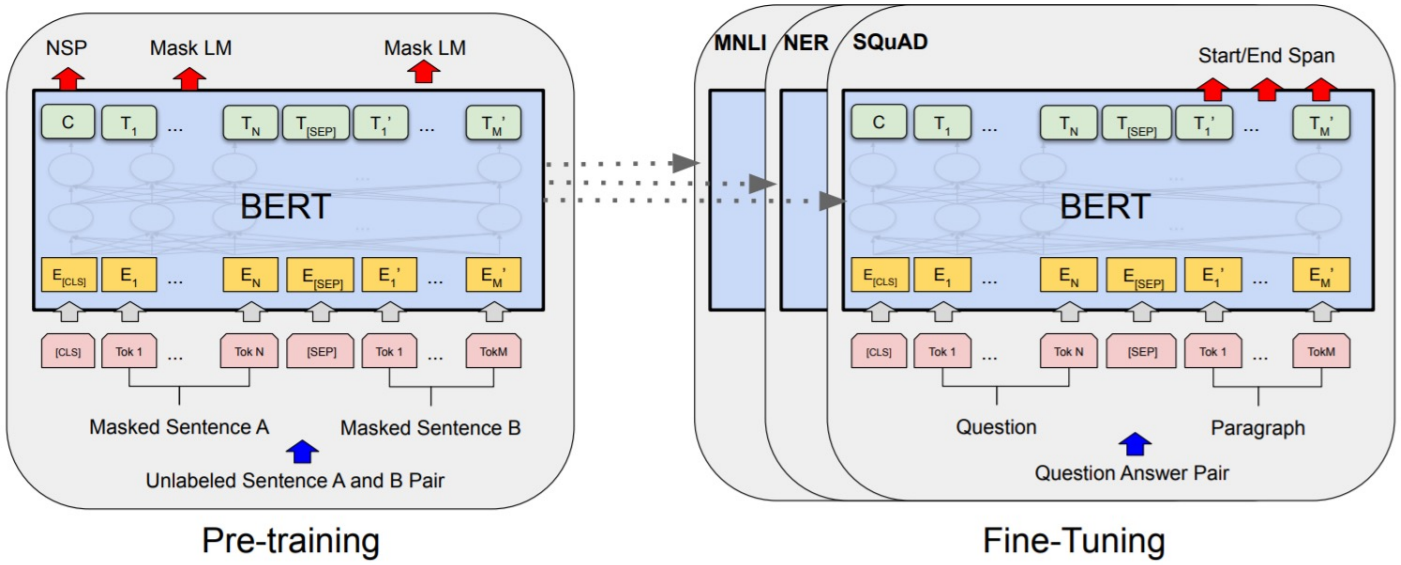


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

GPT-3

- Brown et al., 2020
- Auto-Regressive (Looks to previous context)

Johns Hopkins University was _____

GPT-3

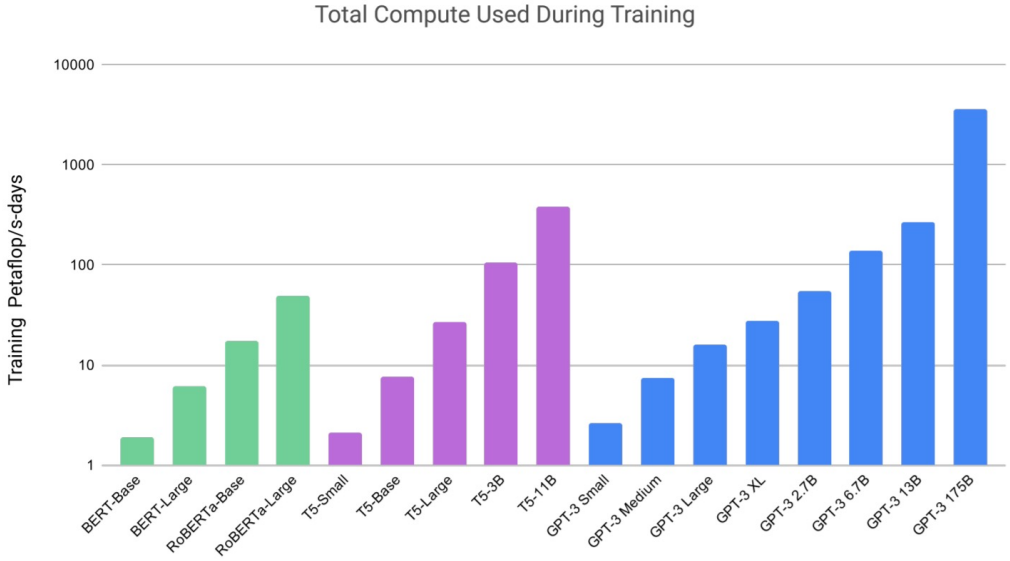


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

GPT-3

- WMT '14 MT
- SacreBLEU (proper eval)
- Trained only on English

Setting	En→Fr
SOTA (Supervised)	**45.6**[a]
XLM [LC19]	33.4
MASS [STQ+19]	37.5
mBART [LGG+20]	-
GPT-3 Zero-Shot	25.2
GPT-3 One-Shot	28.3
GPT-3 Few-Shot	32.6

GPT-4

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

REMINDER: I'M HIRING PHD STUDENTS

- https://docs.google.com/forms/d/e/1FAIpQLScZqqr7xN63lpP7m7Z22_oEVa257f7_jJp4YOBTDURfWck1JA/viewform?usp=sharing&ouid=117382557743098706243