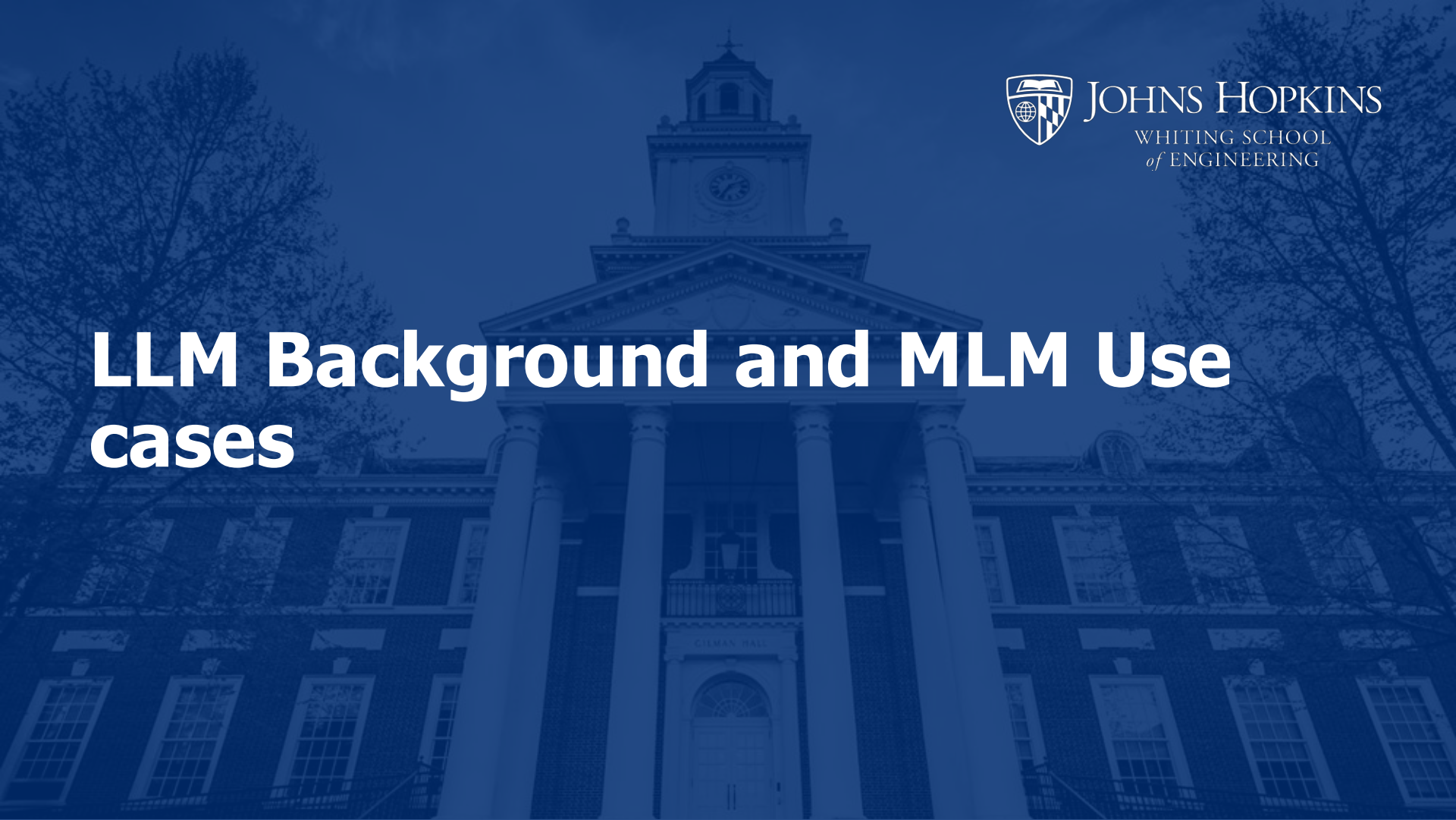
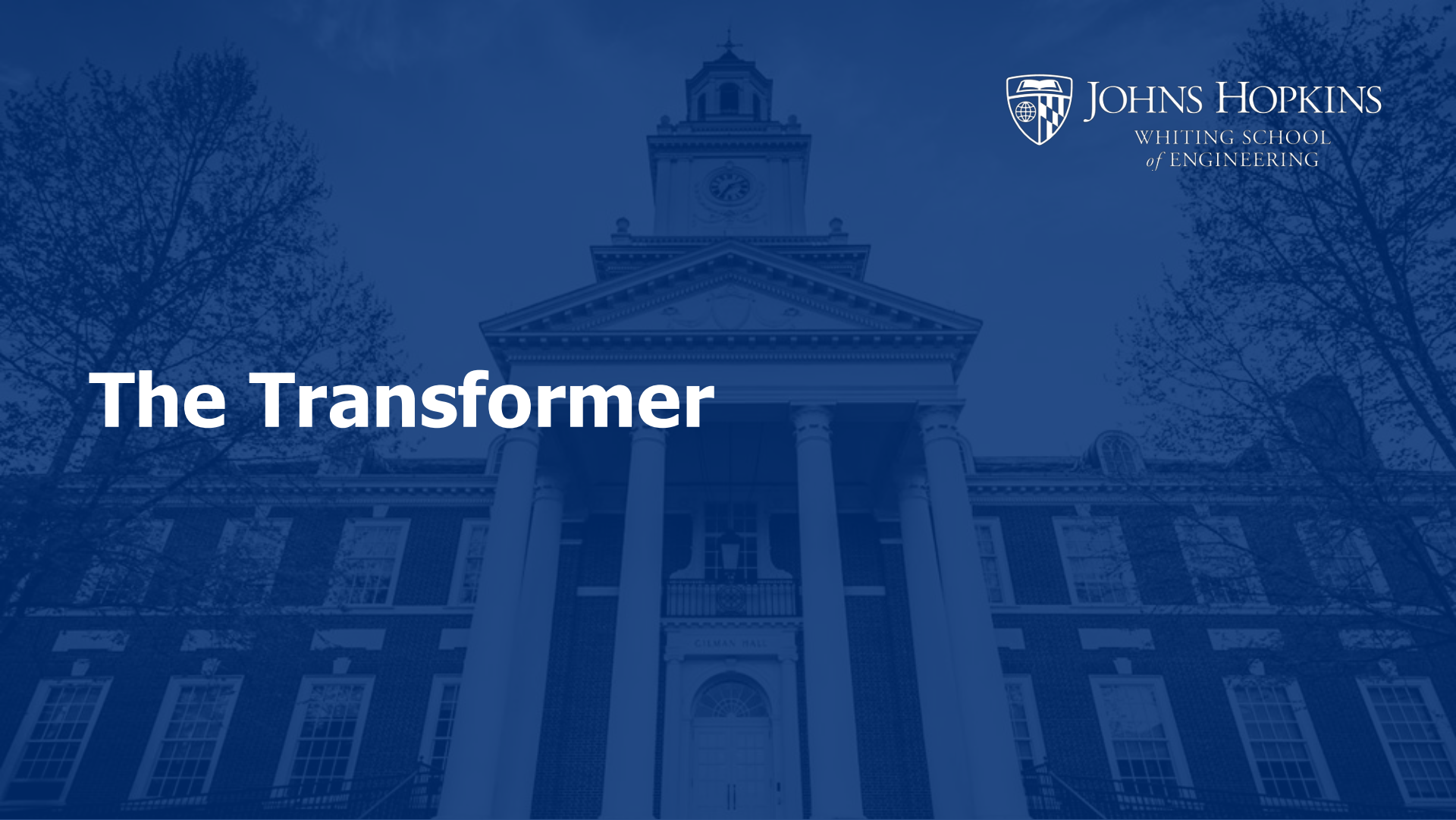


LLM Background and MLM Use cases

Overview

- Last class:
  - N-gram language models, neural models, ELMo
- Brief background on LLMs
  - Transformer
  - Training Objectives
- Social applications of MLMs
  - Polarization and dehumanization
  - Anthropomorphism



The Transformer

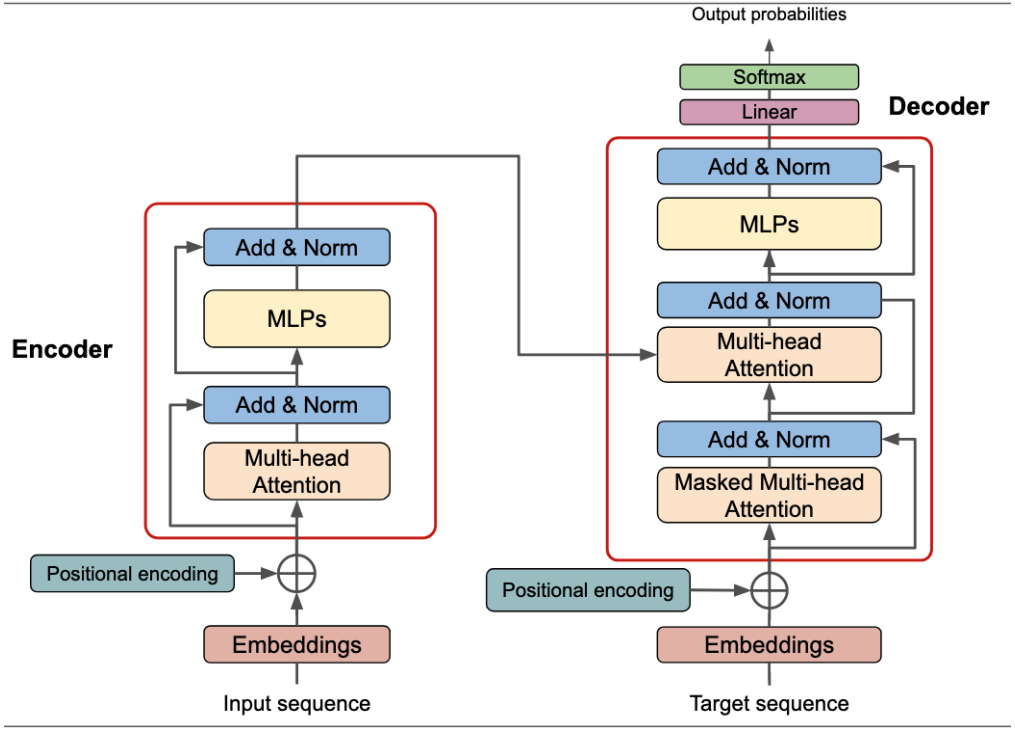
Recap: Predecessor RNNs

- While RNNs in theory can represent long sequences, they quickly forget portions of the input.
- Vanishing/exploding gradients
- Difficult to parallelize
- The alternative architecture: Transformers

"Neural machine translation by jointly learning to align and translate" Bahdanau etl. 2014; "Attention is All You Need" Vaswani et al. 2017 https://arxiv.org/abs/1706.03762]

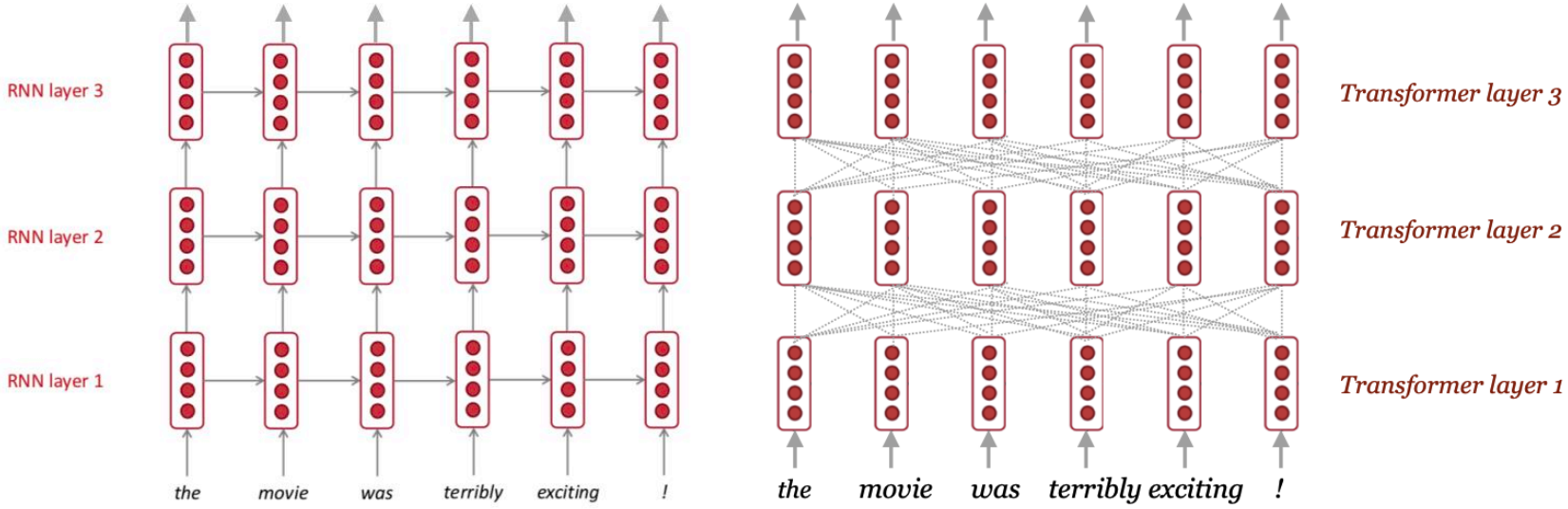
The Transformer

- Stacks of transformer blocks, each of which is a multilayer network that maps sequences of input vectors (x_1,..., x_n) to sequences of output vectors (z_1,..., z_n) of the same length
- Blocks are made by combining simple linear layers, feedforward networks, and self-attention layers (the key innovation of transformers)



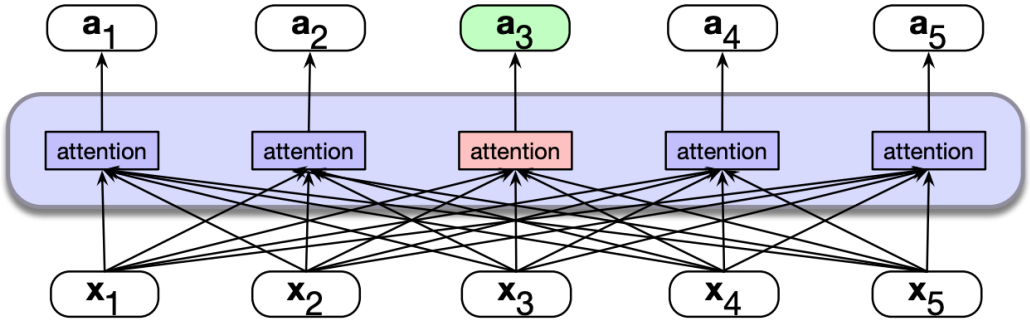
https://deeprevision.github.io/posts/001-transformer/

RNN vs Transformer



Model is able to access all context simultaneously: we need a mechanism to focus ("attend") on a particular part

Defining Attention (simplified)

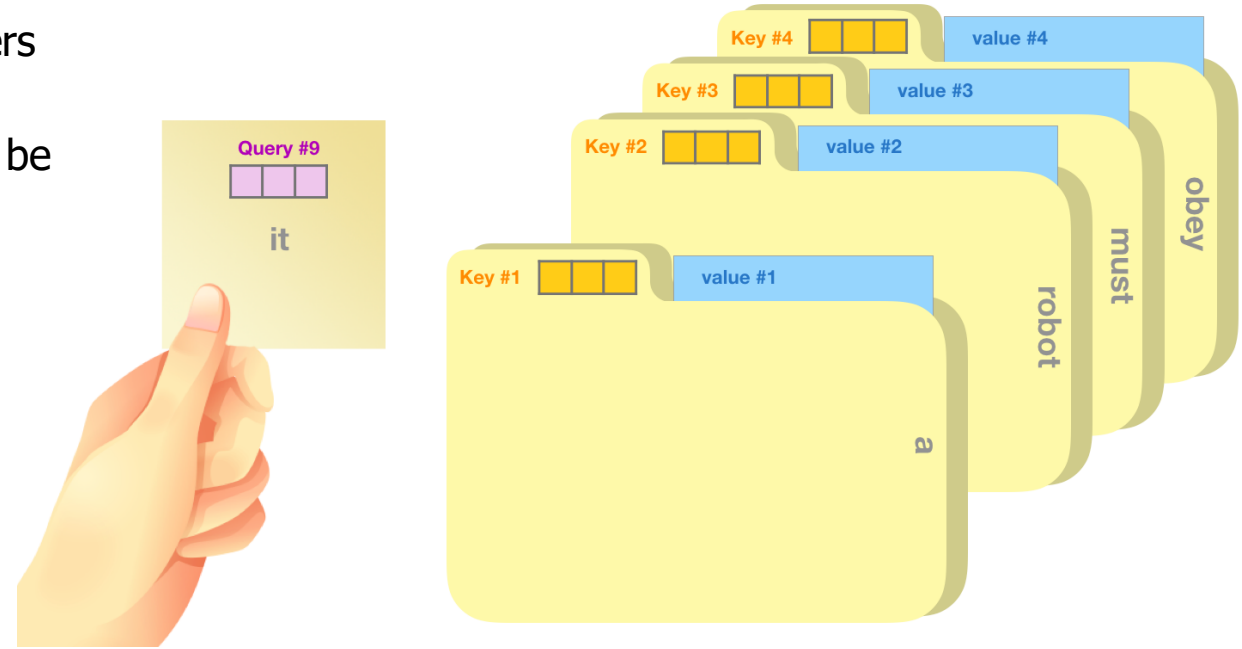


- A weighted sum of vectors

- $a_i = \sum \alpha_{ij} x_j$
  - x_j is the representation for the word j (if this is the first layer)
  - α_{ij} is a scaler used to weight how much x_j should contribute to a_i
  - How do we calculate α_{ij}?

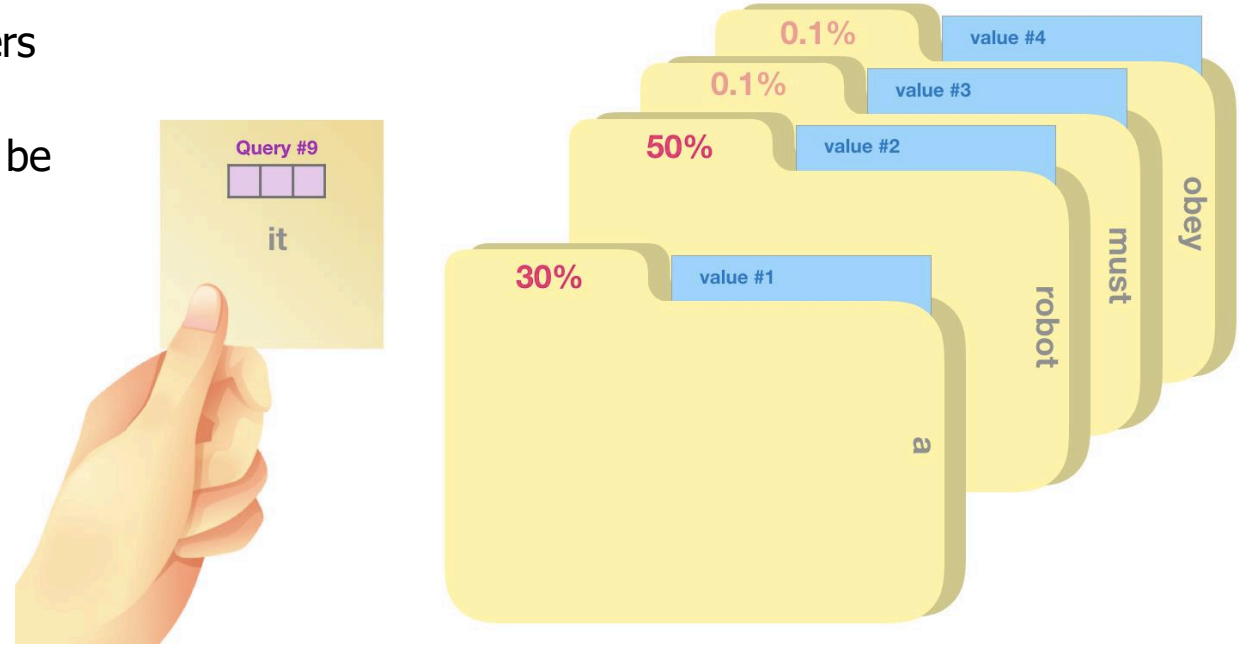
Defining Self-Attention

- Terminology:
  - Query: to match others
  - Key: to be matched
  - Value: information to be extracted



Defining Self-Attention

- Terminology:
  - Query: to match others
  - Key: to be matched
  - Value: information to be extracted



q: query (to match others)

$$q_i = W^q x_i$$

k: key (to be matched)

$$k_i = W^k x_i$$

v: value (information to be extracted)

$$v_i = W^v x_i$$

q: query (to match others)

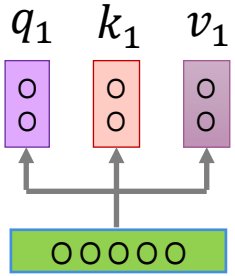
$$q_i = W^q x_i$$

k: key (to be matched)

$$k_i = W^k x_i$$

v: value (information to be extracted)

$$v_i = W^v x_i$$



The

q: query (to match others)

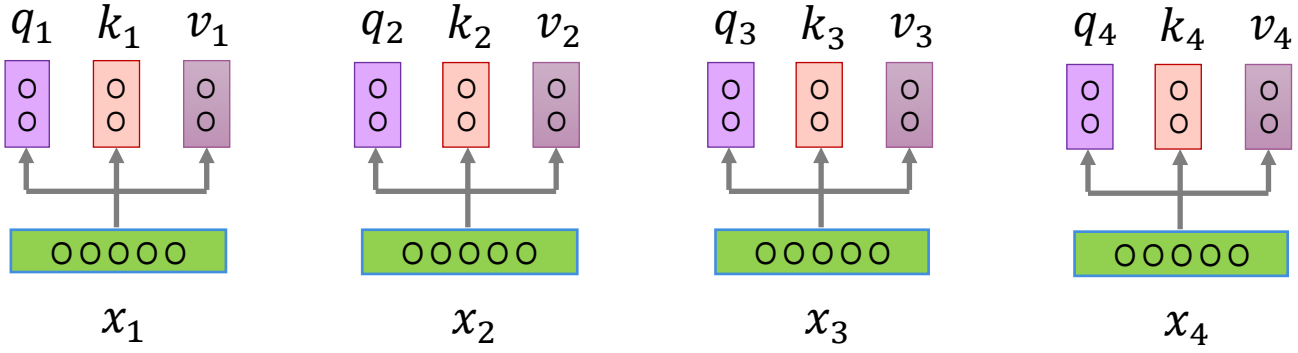
$$q_i = W^q x_i$$

k: key (to be matched)

$$k_i = W^k x_i$$

v: value (information to be extracted)

$$v_i = W^v x_i$$



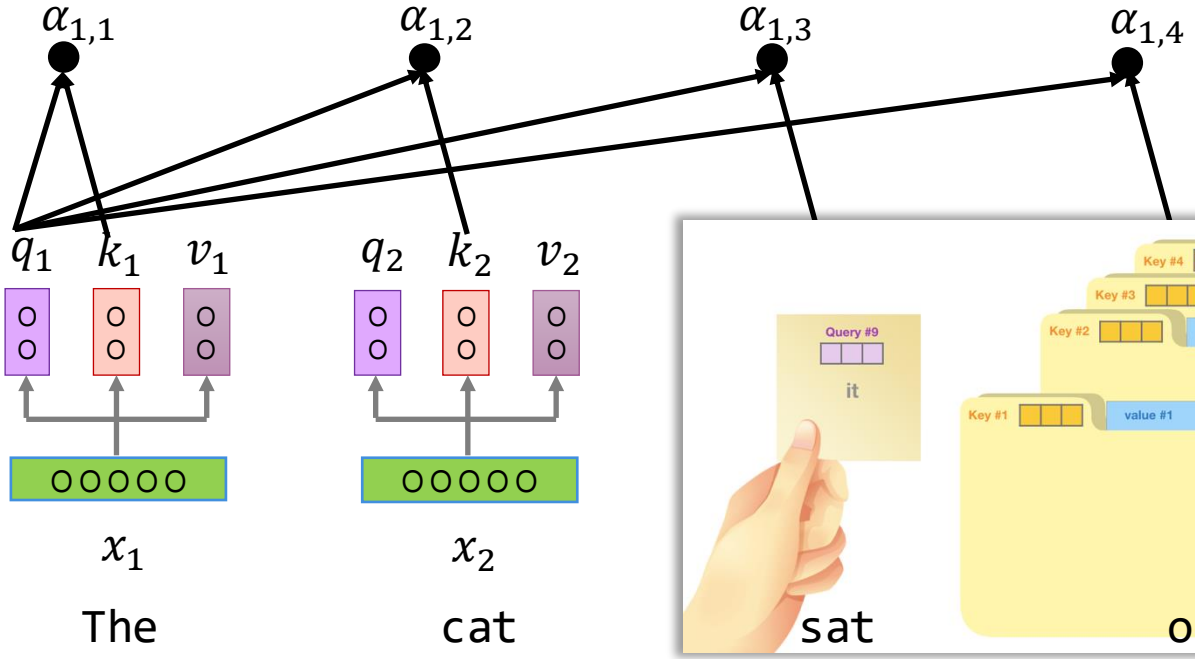
$$\alpha_{1,i} = \underbrace{q^1 \cdot k^i / \sqrt{d}}_{\text{Scaled dot product}}$$

How much should "The" attend to other positions?

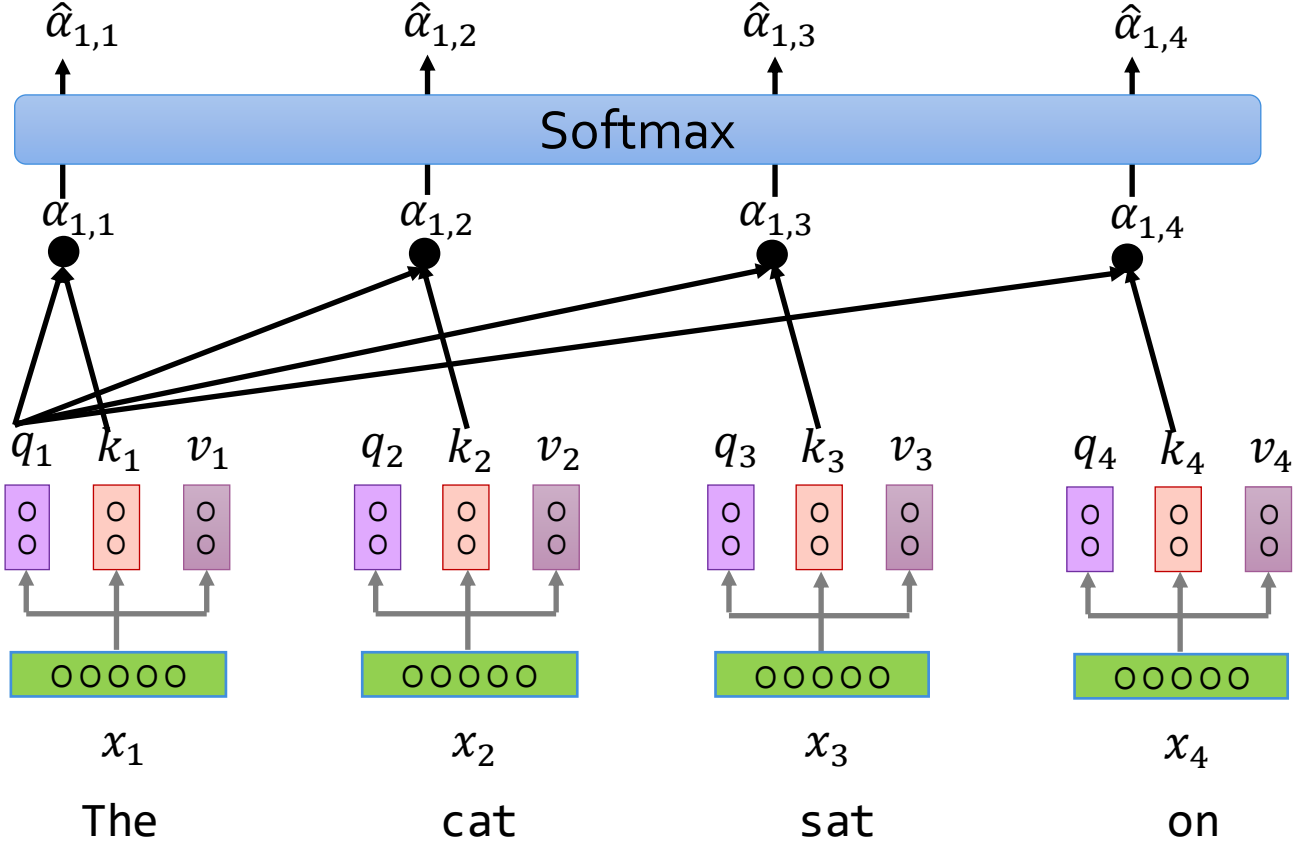
q: query (to match others)

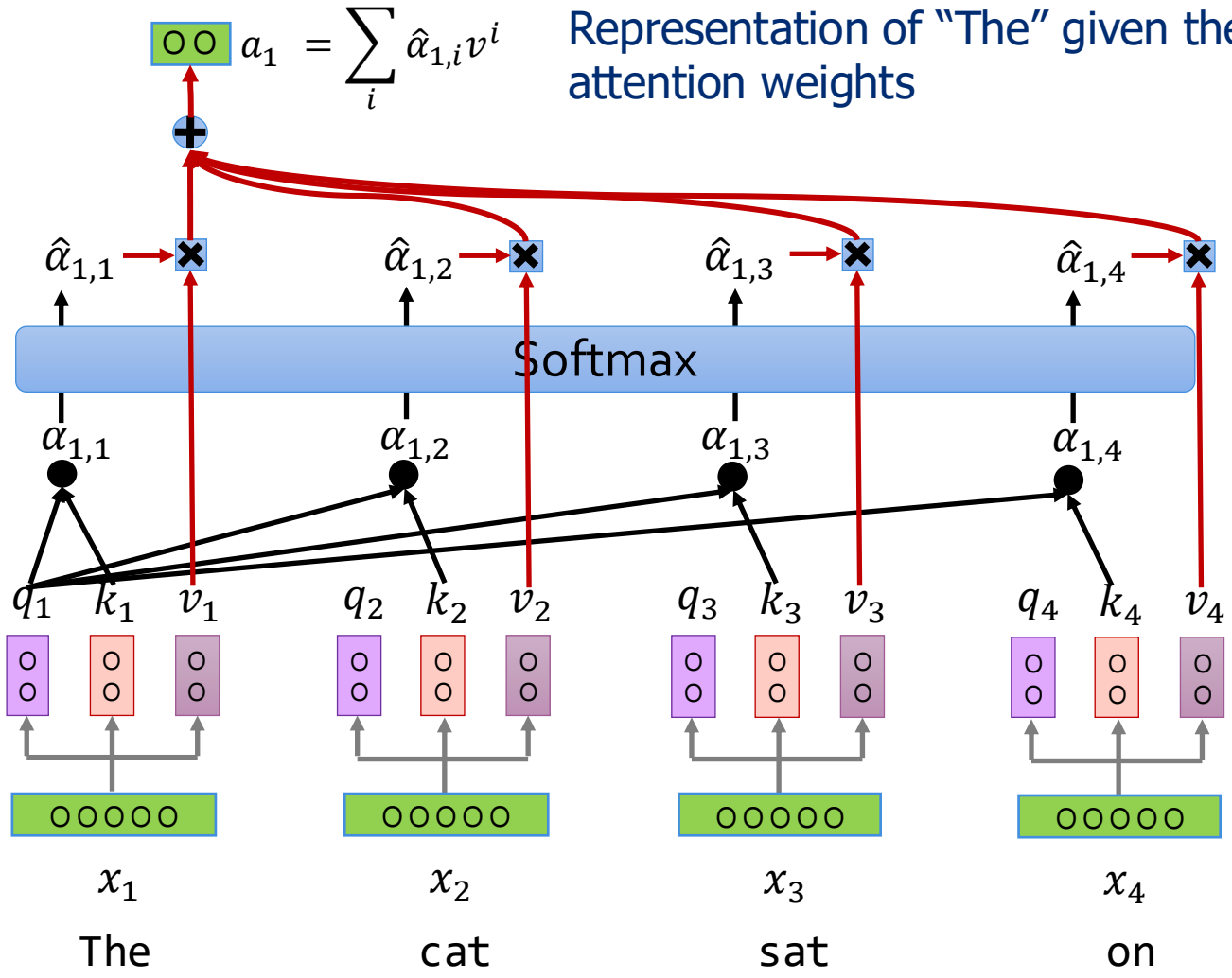
k: key (to be matched)

v: value (information to be extracted)



$$\sigma(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$




$a_1 = \sum_i \hat{\alpha}_{1,i} v^i$
Representation of "The" given the attention weights
$\hat{\alpha}_{1,1}$
$\hat{\alpha}_{1,2}$
$\hat{\alpha}_{1,3}$
$\hat{\alpha}_{1,4}$
Softmax
$\alpha_{1,1}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,4}$
q_1 k_1 v_1
q_2 k_2 v_2
q_3 k_3 v_3
q_4 k_4 v_4
x_1
x_2
x_3
x_4
The
cat
sat
on

Self-Attention: Matrix Notation

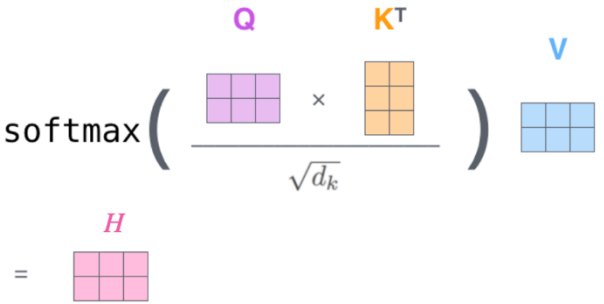
$X \in \mathbb{R}^{n \times d_1}$ (n = input length)

$Q = XW^Q \quad K = XW^K \quad V = XW^V$

$W^Q \in \mathbb{R}^{d_1 \times d_q}, W^K \in \mathbb{R}^{d_1 \times d_k}, W^V \in \mathbb{R}^{d_1 \times d_v}$

$n \times d_q$ $d_k \times n$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$ $n \times$



[Slide credit: Danqi Chen]

Properties of Self-Attention

Layer Type	Complexity per Layer	Sequential Operations
Self-Attention	$O(n^2 \cdot d)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$

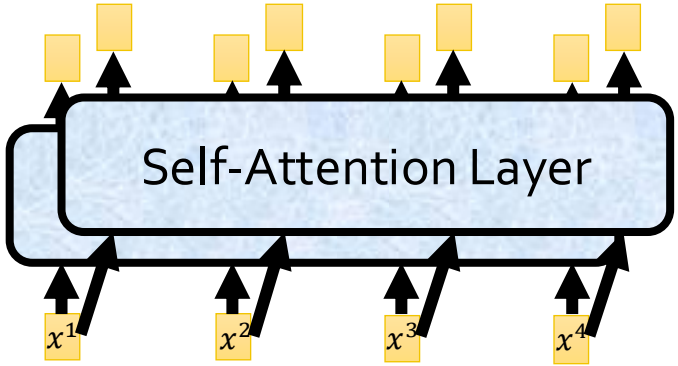
- n = sequence length, d = hidden dimension
- Quadratic complexity, but:
  - O(1) sequential operations (not linear like in RNN)

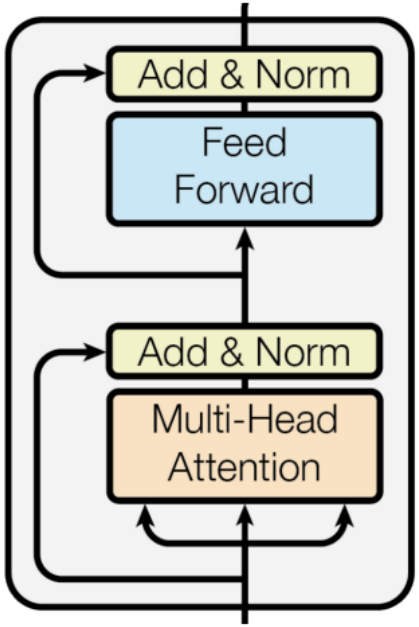
- Efficient implementations

[Attention Is All You Need, Vaswani et al. 2017]

Additional components of transformers

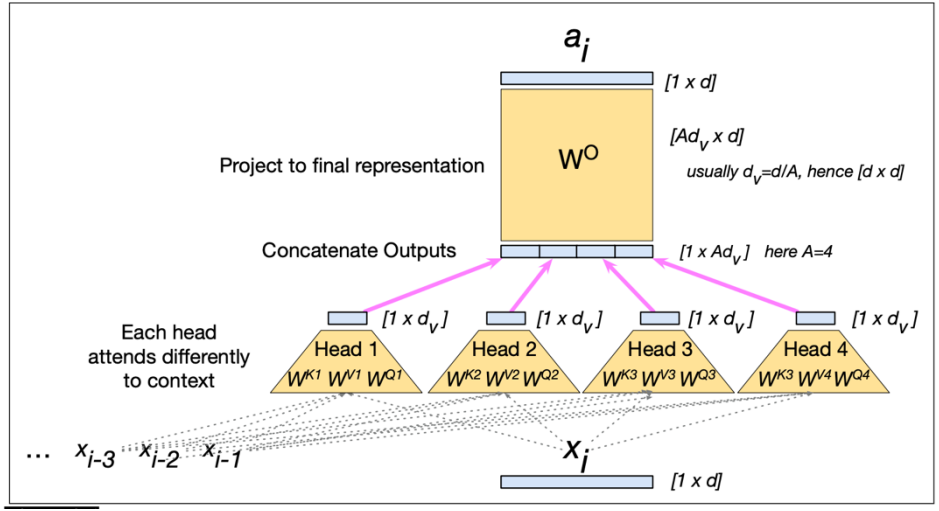
- Multi-head self-attention: multiple parallel attention layers
  - Each attention layer has its own parameters.
  - Concatenate the results and run them through a linear projection.

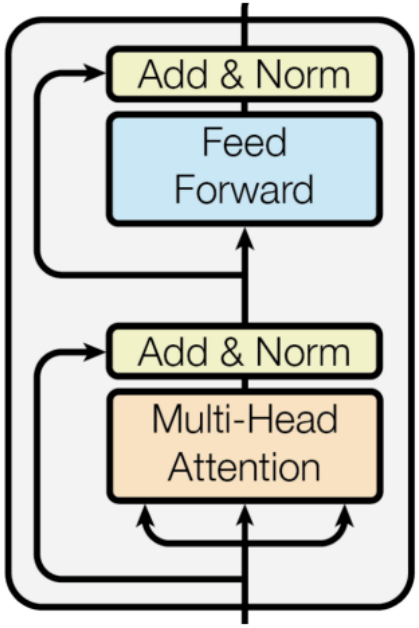




Additional components of transformers

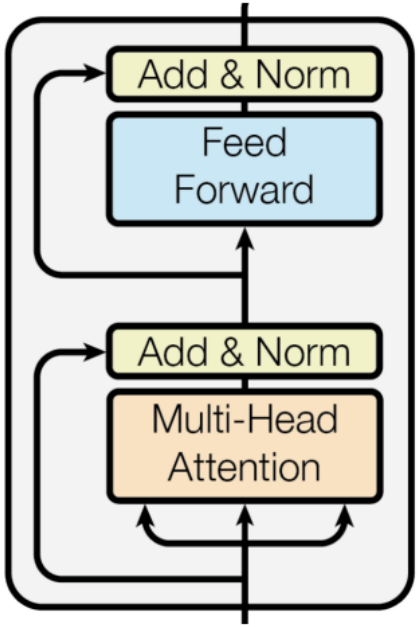
- Multi-head self-attention: multiple parallel attention layers
  - Each attention layer has its own parameters.
  - Concatenate the results and run them through a linear projection.





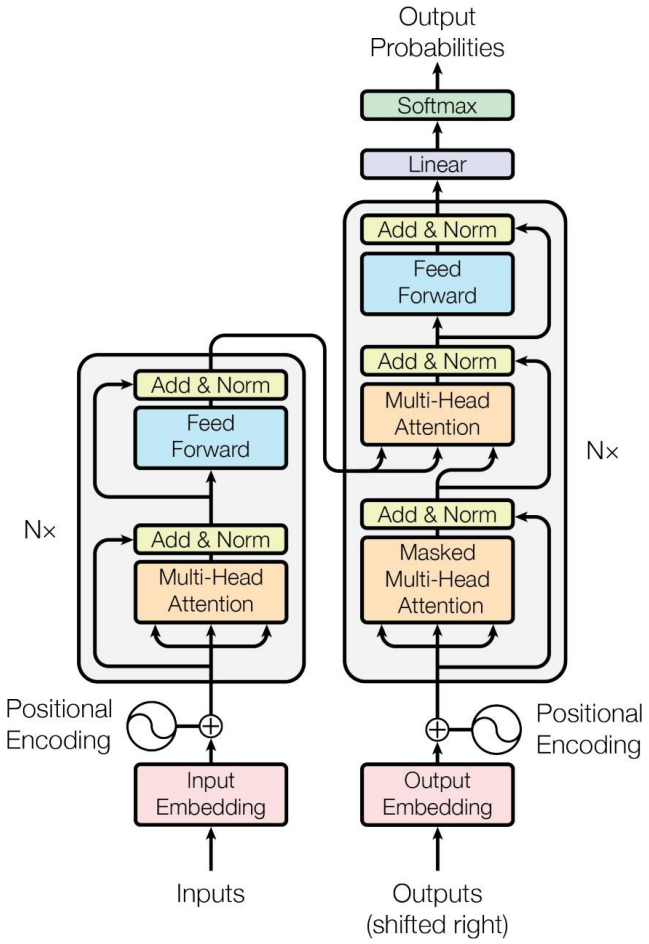
Additional components of transformers

- Multi-head self-attention: multiple parallel attention layers
- **Positional embeddings**: we've lost the notion of word order and need embeddings to specifically track it
- **Residual connections** let the model "skip" layers
- **Layer normalization**
- **Feed-forward layers**



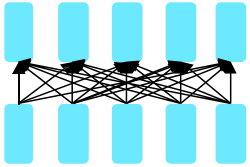
Encoder-decoder Models

- The original transformer architecture was encoder decoder
- Decoder attends to previous computation of encoder as well as decoder's own generations
- Encoder-decoder models are flexible in both generation and classification tasks



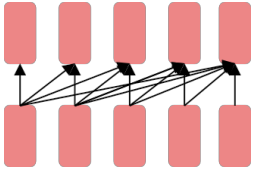
Variants of Transformers

- A building block for a variety of LMs



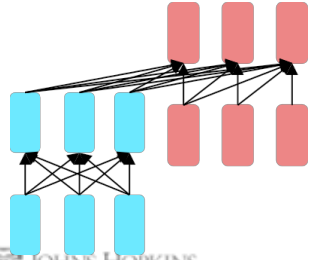
Encoders

- Examples: BERT, RoBERTa, SciBERT.
- Captures bidirectional context?



Decoders

- Examples: GPT-2, GPT-3, LaMDA
- Other name: causal or auto-regressive language model
- Nice to generate from; can't condition on future words



Encoder-Decoders

- Examples: Transformer, T5, Meena

Training Objectives

Language model training objectives

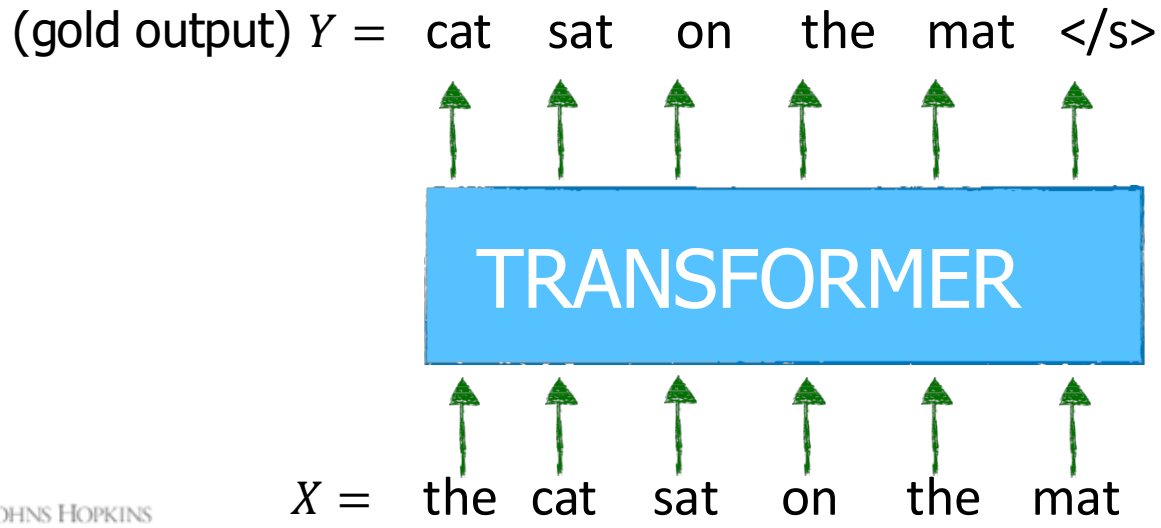
- Recall: we need self-supervised training objective

- Two common approaches:
  - Next token prediction (e.g. GPT models)
  - Masked language modeling (e.g. BERT, RoBERTa, and other follow up variants)

Next Token Prediction

- **Goal:** Train a Transformer for language modeling (i.e., predicting the next word).
- **Approach:** Train it so that each position is predictor of the next (right) token.
  - We just shift the input to right by one, and use as labels

EOS special token

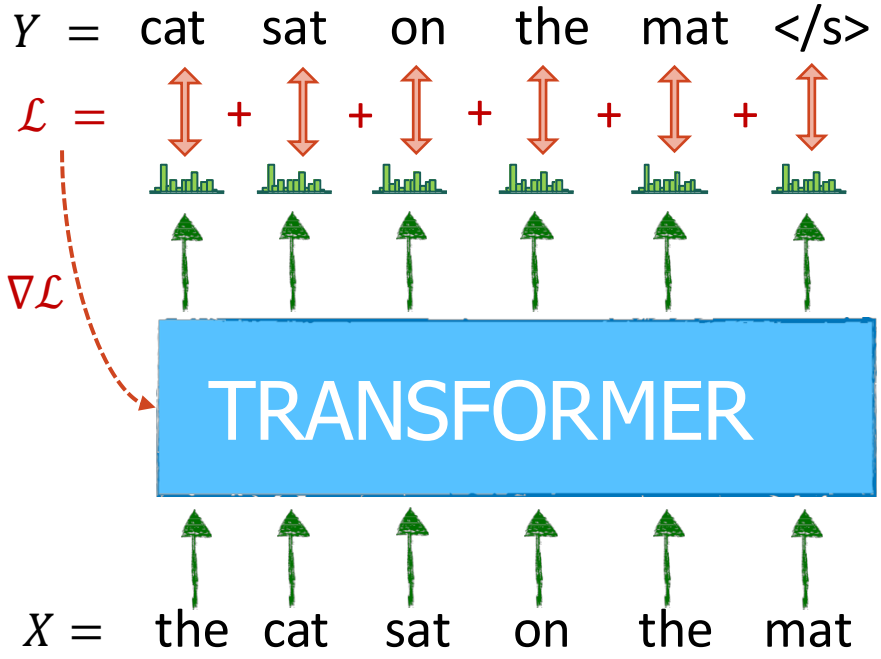



```
X = text[:, :-1]
Y = text[:, 1:]
```

[Slide credit: Arman Cohan]

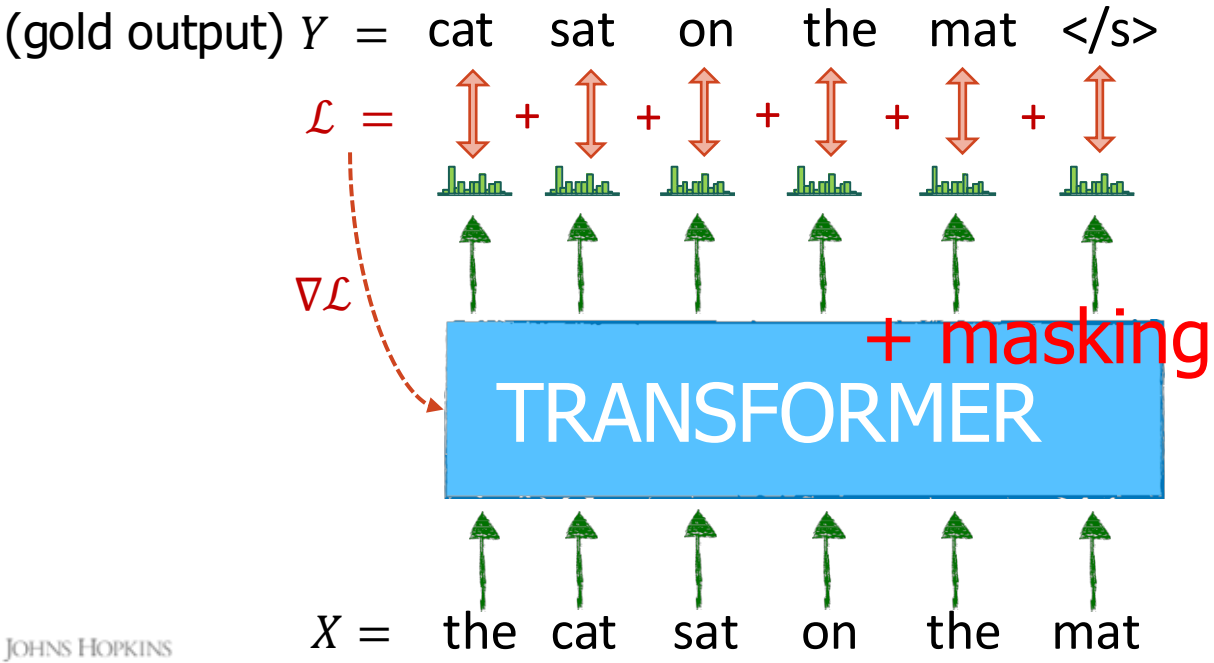
Training a Transformer Language Model

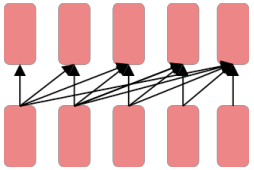
- The model would solve the task by copying the next token to output (data leakage).
  - Does not learn anything useful



Training a Transformer Language Model

- We need to **prevent information leakage** from future tokens



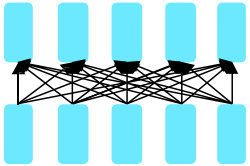


Alternative pretraining objective: masking

the man went to the ~~store~~ to buy a ~~gallon~~ of milk

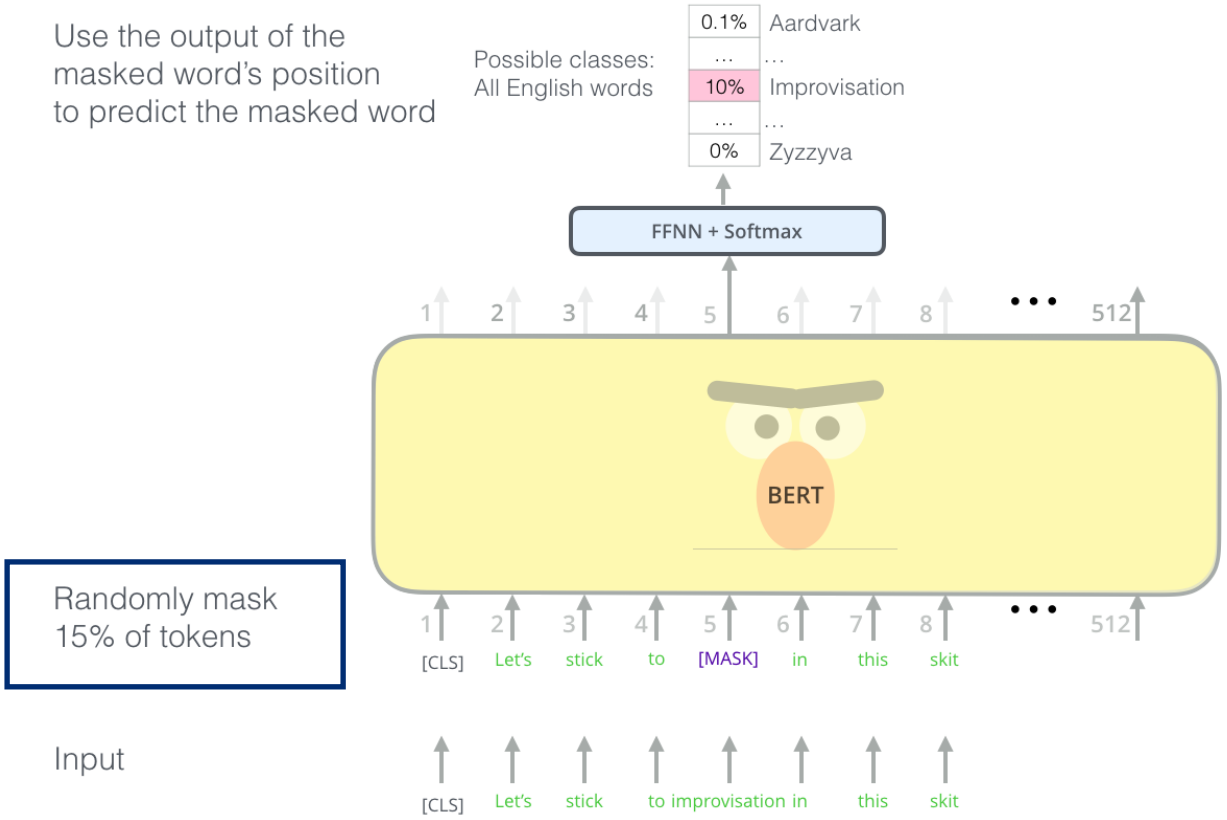
the man went to the [MASK] to buy a [MASK] of milk

This is the pre-training objective used for Bidirectional Encoder Representations from Transformers (**BERT**) and similar encoder-only models



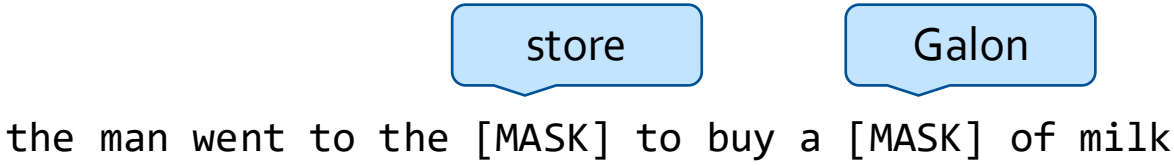
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (1): Masked Tokens



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (1): Masked Tokens



- Too little masking: Too expensive to train
- Too much masking: Underdefined
  - (not enough info for the model to recover the masked tokens)

Later work shows that more principled masking (instead of uniformly random) could benefit downstream task performance and result in faster training.

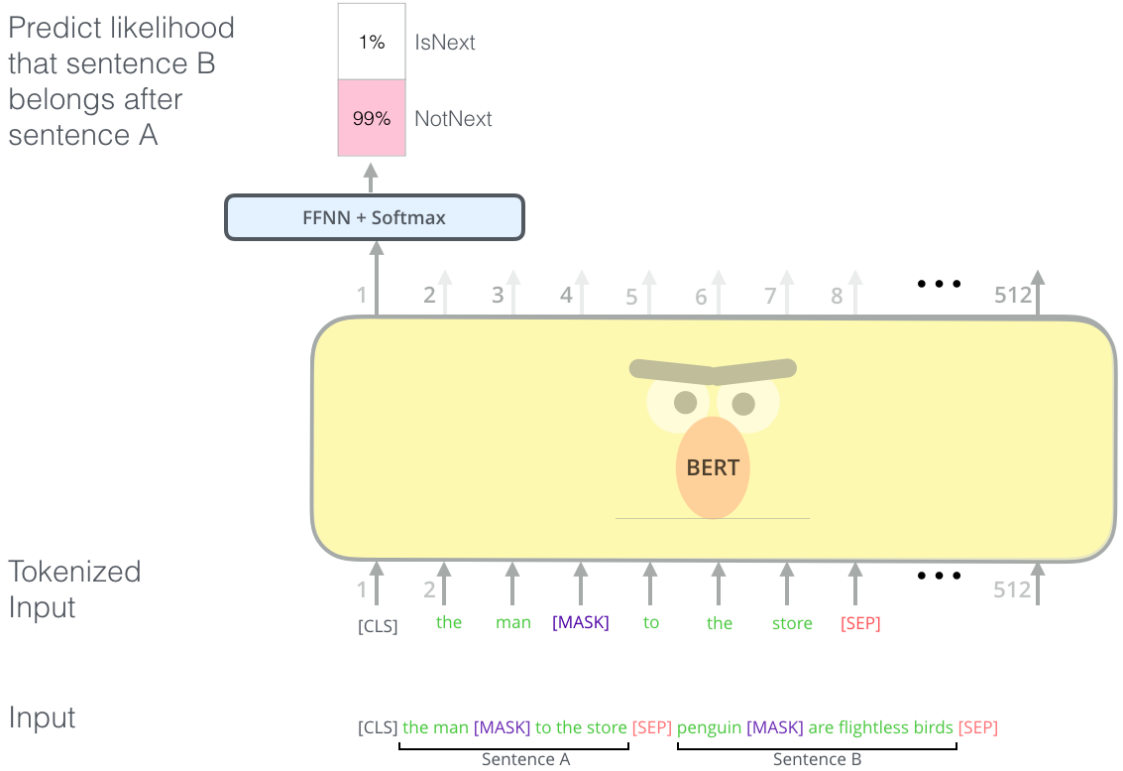
PMI Masking (Levine et al., 2021) https://arxiv.org/pdf/2010.01825.pdf
SpanBERT (Joshi et al., 2020) https://arxiv.org/pdf/1907.10529.pdf

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (2): Sentence Ordering

- Predict sentence ordering

- 50% correct ordering, and 50% random incorrect ones



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

MLM example use cases: measuring dehumanization and anthropomorphism

How can we use pre-trained language models?

- Powerful classifiers for supervised tasks
  - Need less data than training a model from scratch
- What about more open-ended text analysis tasks?
  - Can we think of creative use cases for these models?
- Think about:
  - Can you think of other scenarios where this use of MLMs might be useful?
  - How might you improve the evaluation conducted in these projects?
  - What are some of the limitations? How do they limit conclusions?

Motivation

- Rise of vocal anti-immigrant politicians in the US in recent years
  - Anecdotally, seems that attitudes towards immigration have become more negative (or at least more polarized)
- Historically, resistance to newcomers has always been a central part of US public discourse about immigration
  - Anti-Chinese fearmongering in the 1880s
  - Concerns about Southern and Eastern European immigrants in the 1920s
- Move beyond anecdotes:
  - How have attitudes toward immigrants in the United States changed over the past century?
  - How does recent political debate over immigration compare to the long sweep of US history?

Card, Dallas, et al. "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration." Proceedings of the National Academy of Sciences 119.31 (2022): e2120510119.

Data

- Challenge:
  - Public opinion polls that asked about attitudes toward immigration only began in the 1960s
  - Until recent years, polls only asked about immigration sporadically

- Instead:
  - Analyze discussions of immigration in free-form text data
  - 7 million congressional speeches from 1880 to the present
  - ~200,000 speeches relevant to the topic of immigration

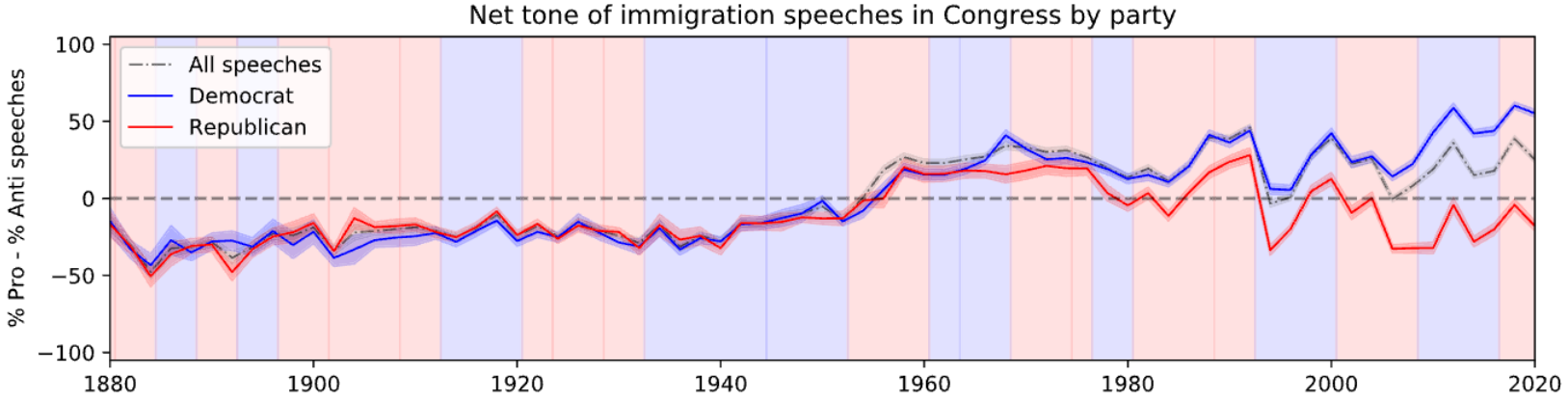
Analysis 1: Tone (Methods)

- Tone (pro-immigration, anti-immigration or neutral)

- Data annotations:
  - Hand-label congressional speeches as about immigration or not; hand-label tone
  - At least 2 annotators per segment

- Model:
  - Base model: RoBERTa
  - Tune the model to the data set using self-supervised training over congressional speeches
  - Tune the model to classify relevance (90% accuracy) and tone (65% accuracy)
  - "vast majority of tone errors are between neutral and one of the extremes"

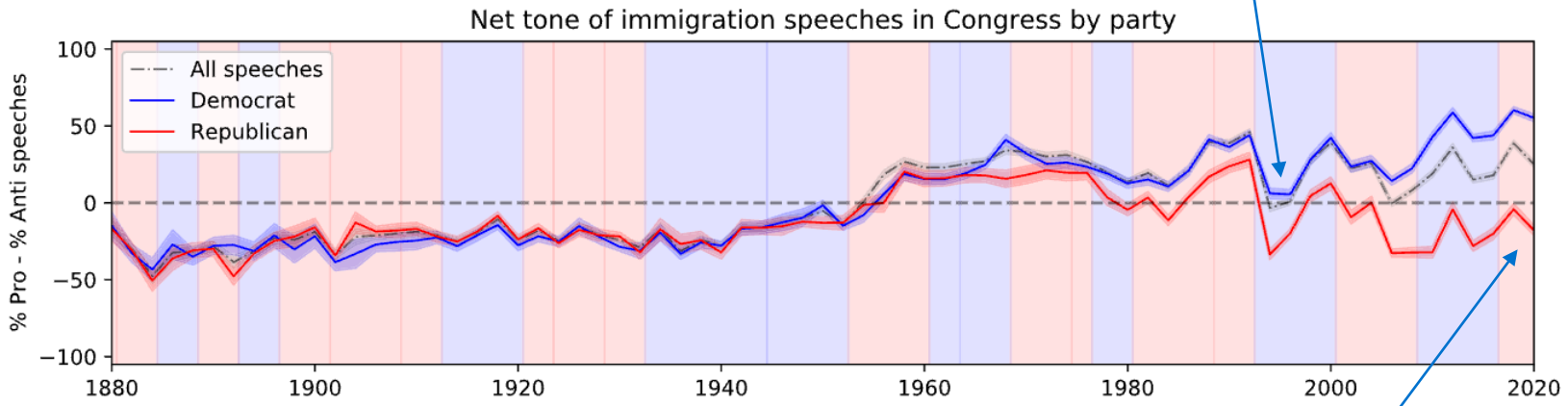
Analysis 1: Tone



- Tone in Congress has been positive since about WW2
- Divergence in tone between Democrats and Republicans

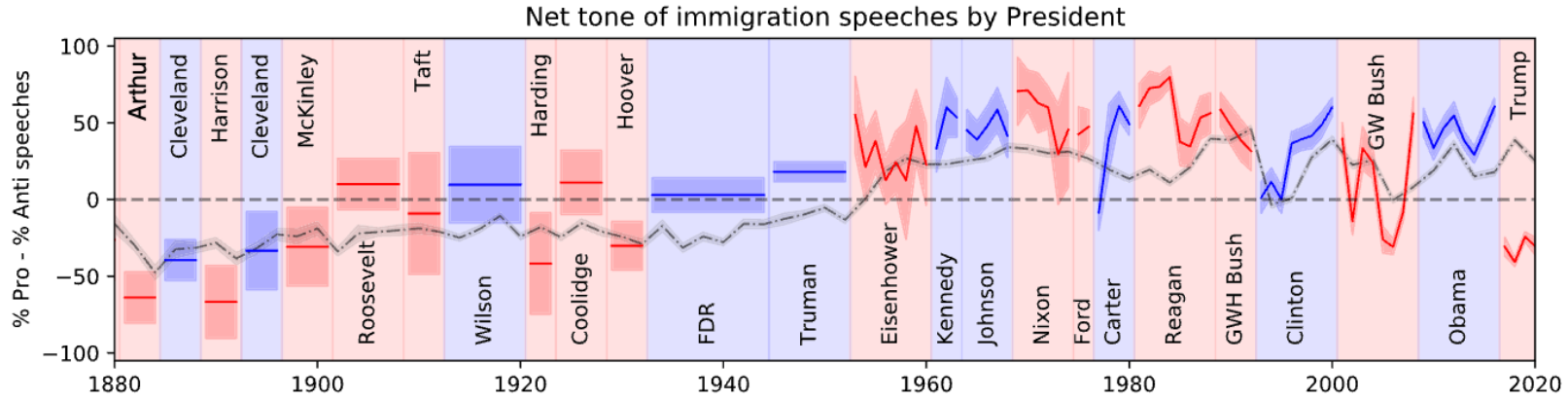
Analysis 1: Tone

Democrats, have grown more positive about immigration over time, "with the exception of a temporary bipartisan drop in pro-immigration speeches in the early 1990s, coinciding with the end of the Cold War and the passage of the North American Free Trade Agreement (NAFTA)"



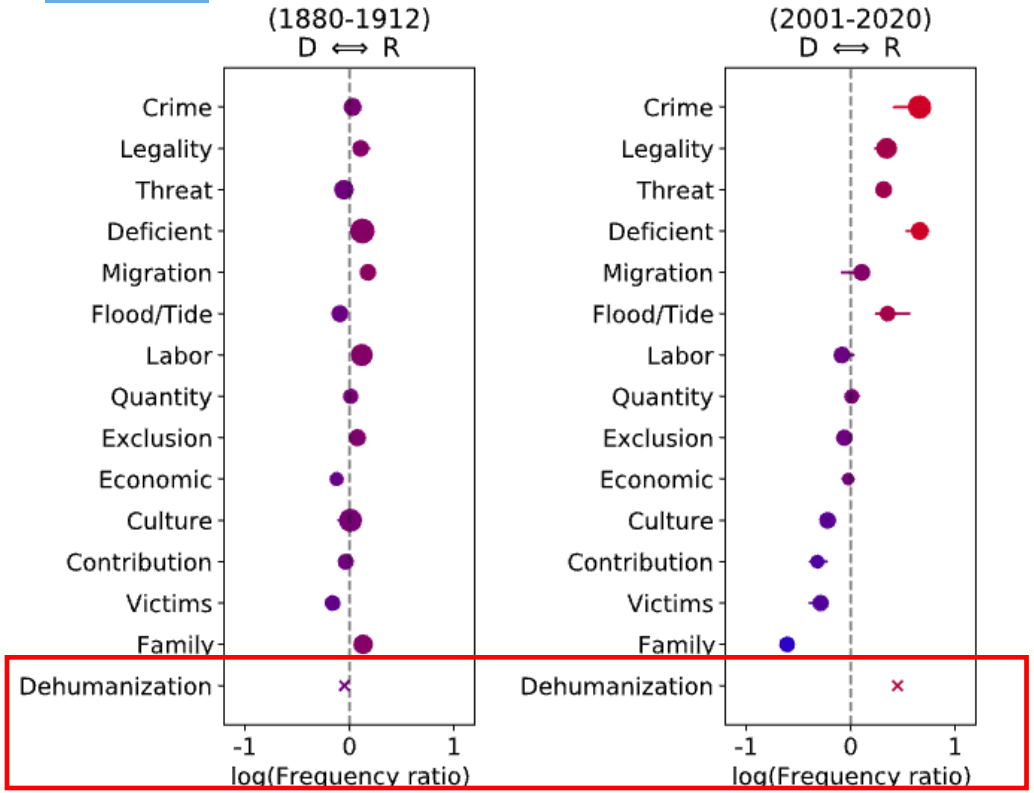
Republican legislators are now "approximately as overtly antiimmigration in their speeches as the average legislator was during the Age of Mass Migration from Europe and the 1920s quota periods"

Analysis 1: Tone (Presidential Speech)



- Tone in presidential speeches has also been almost entirely positive, especially since WW2
- Trump is stark exception

More nuanced language analysis



- Log frequency ratio of terms in framing lexicons (curated automatically and manually)

Measurement of Dehumanization

- Observation: we can leverage the way masked language models were trained to identify *metaphorical* language

- Example: identify *dehumanizing language*
  - Original: "The children scurried away"
  - Ask model to predict: "The _____ scurried away"
  - Model predicts non-human word (e.g. "*mice*") with high probability → the sentence using dehumanizing language

- More harmful: "The illegal immigrants scurried over the border"

Measurement of Dehumanization

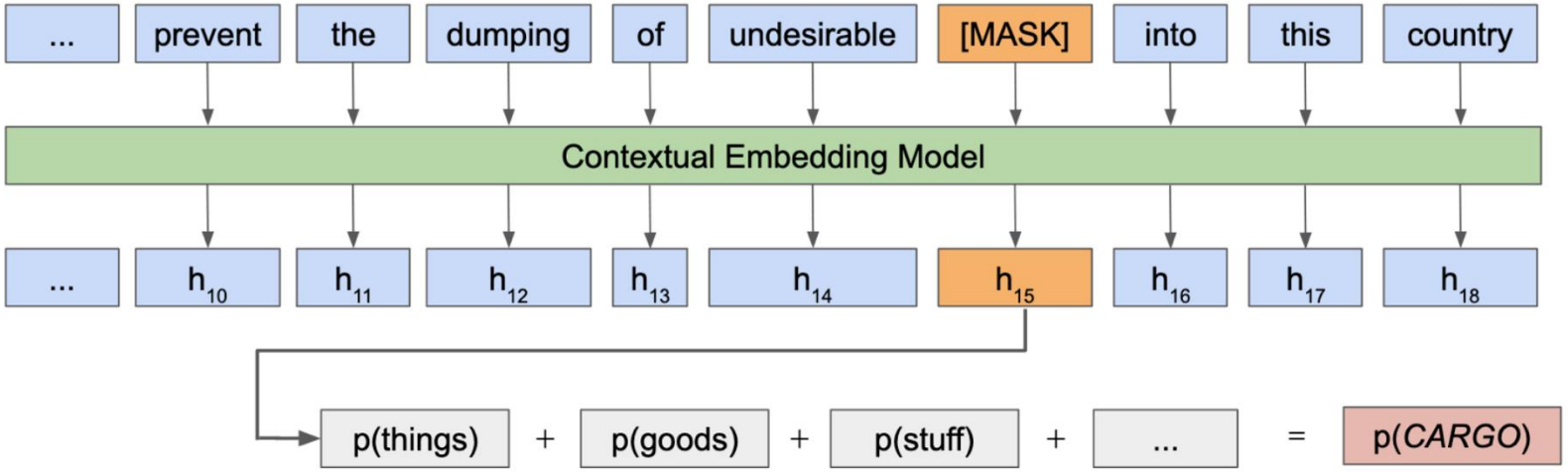
- Build lexicons of relevant words
  - Starting point: Metaphors from prior work about immigration and dehumanization, such as “animals” and “cargo”
  - Curate initial seed list of terms (e.g. animal, animals, etc)
  - Use static word embeddings to identify similar words and add them to the list (e.g. automated lexicon expansion)

Metaphor	Terms in BERT vocabulary
Animals	animal, animals, beast, beasts, brute, cattle, cow, cows, dog, dogs, herd, herds, hog, horse, horses, livestock, pig, pigs, sheep
Cargo	thing, things, object, objects, cargo, goods, merchandise, item, items, commodities, packages, products, baggage, shipment, shipments, stuff, material
Disease	disease, diseases, virus, viruses, infection, infections, illness, illnesses
Flood/Tide	flood, flooding, floods, ocean, oceans, river, rivers, stream, tide, tides, water, waters, wave, waves
Machines	machine, machines, machinery, equipment, apparatus, appliances, hardware, engine, engines, tool, tools, device, devices
Vermin/Pests	rat, rats, worm, worms, bug, bugs, parasite, parasites, insect, insects, pest, flea, rodents
Random	adoption, aerial, agricultural, amtrak, announcements, antenna, brave, cadet, captures, carroll, champaign, charley, ecosystem, excuses, exit, french, freshman, goal, headache, inter, knock, liberty, lifeboat, london, manifest, mrs, multimedia, narcotics, nitrate, orr, ow, parliamentary, plantation, proof, protect, provider, ready, reese, revolutionaries, ribbons, san, sanders, satisfaction, scope, series, sucker, superstructure, whig, whiskey

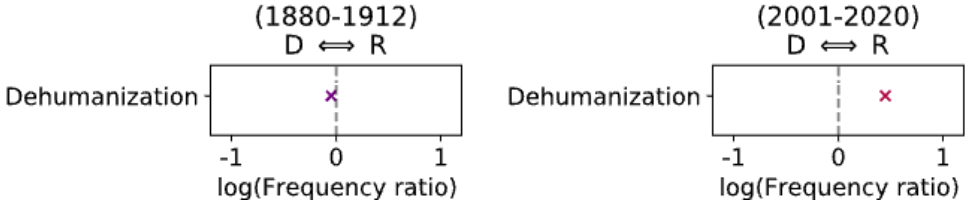
Measurement of Dehumanization

- For each sentence that mentions an immigrant or immigrant group:
  - Replace the mention (e.g., "foreigners") with BERT's special "[MASK]" token—indicating a gap to be filled
  - Process the masked sentences through the model and compute how likely it is (according to the model) that the gap would be filled by each term in each metaphorical category
  - Add up the probabilities for each term in the category to get a score for the entire category

Measurement of Dehumanization



Measurement of Dehumanization



- Higher frequency of dehumanizing metaphors in Republican Speech after 2001

0.87	1939	aliens	the destruction of these homes by a ruthless government . cruel separation of families . and the herding of these [MASK] in stockades is pictured .
0.82	1963	Mexican nationals	it was enacted at that time in order to provide effective control procedures for the movement of [MASK] into the farmlands in the united states .
0.80	1947	Cuban men	it happened to b - general weyler . who was herding [MASK] . women . and children in concentrados .

- Examples of sentences with high detection of "animal" metaphor

Validation

- Collect human judgements on a sample of masked contexts
- Three of the authors independently rated whether an animal term would be a plausible replacement for the mask token, given the surrounding context
- Annotations showed reasonably strong agreement (Krippendorff's alpha=0.59) and correlated strongly with the log probabilities assigned by the model (r=0.73)

Motivation

- Anecdote:
  - We're increasingly using anthropomorphism when talking about technology
    - "The model learns to do X"
    - "The neural network figures out what features are important"

- Why does this matter?
  - "Projecting human qualities onto these tools facilitates misinformation about their true capabilities, over-reliance on technology, and corporate avoidance of responsibility"

- [Note that anthropomorphism is not inherently harmful]

Myra Cheng, et al. "AnthroScore: A Computational Linguistic Measure of Anthropomorphism" EACL (2024).

Methodology

- Goal: Measure anthropomorphism
- Observation: "humanization" is the opposite of "dehumanization" → we can use a really similar method as in the previous paper!
- Instead of curated lexicons use pronouns: "he", "she" for human and "it" for non-human

$$P_{\text{HUMAN}}(s_x) = \sum_{w \in \text{human pronouns}} P(w),$$

$$P_{\text{NON-HUMAN}}(s_x) = \sum_{w \in \text{non-human pronouns}} P(w),$$

$$A(s_x) = \log \frac{P_{\text{HUMAN}}(s_x)}{P_{\text{NON-HUMAN}}(s_x)}$$

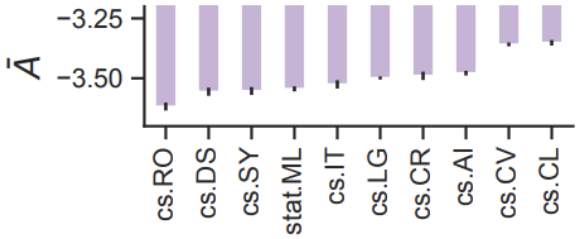
Data

- Abstracts from ~600K papers on CS/Stat arXiv
- ~55K papers in the Association of Computational Linguistics (ACL) Anthology
- Headlines from ~14K downstream news articles that cite these papers.

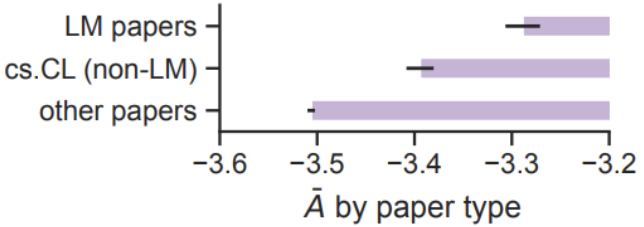
Examples

$S_{\uparrow}$: Sentences with high ANTHROSCORE ($A > 1$)	$S_{\downarrow}$: Sentences with low ANTHROSCORE ($A < -1$)
• When a job arrives, **the system** must decide whether to admit it or reject it, and if admitted, in which server to schedule the job. • Meanwhile, anti-forensic attacks have been developed to fool **these CNN-based forensic algorithms**. • **The models** demonstrated qualifications in various computer-related fields, such as cloud and virtualization, business analytics, cybersecurity, network setup...	• More and more users and developers are using **Issue Tracking Systems** to report issues, including bugs, feature requests, enhancement suggestions, etc. • **Our approach** delivers forecast improvements over a competitive benchmark and we discover evidence for strong spatial interactions. • To this end, for training **the model**, we convert the knowledge graph triples into reasonable and unreasonable texts.
• ***Large language models*** *don't actually think and tend to make elementary mistakes, even make things up.* • ***The algorithms*** *also picked up on racial biases linking Black people to weapons.* • ***The AI system*** *was able to defeat human players in...*	• *Microsoft is betting heavily on integrating* ***OpenAI's GPT language models*** *into its products to compete with Google.* • *Deepmind has been the pioneer in making* ***AI models*** *that have the capability to mimic a human's cognitive...* • *For workers who use* ***machine-learning models*** *to help them make decisions, knowing when to...*

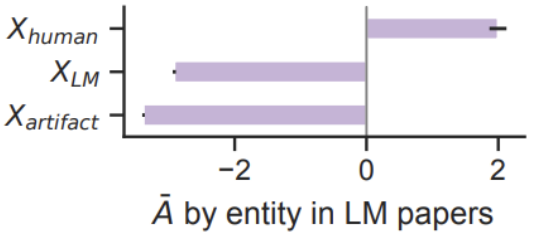
Results: Anthropomorphism is most prevalent in paper abstracts about computational linguistics, and language models



Among the top 10 categories in CS/Stat arXiv, Computation and Language (cs.CL) has the highest average ANTHROSCORE

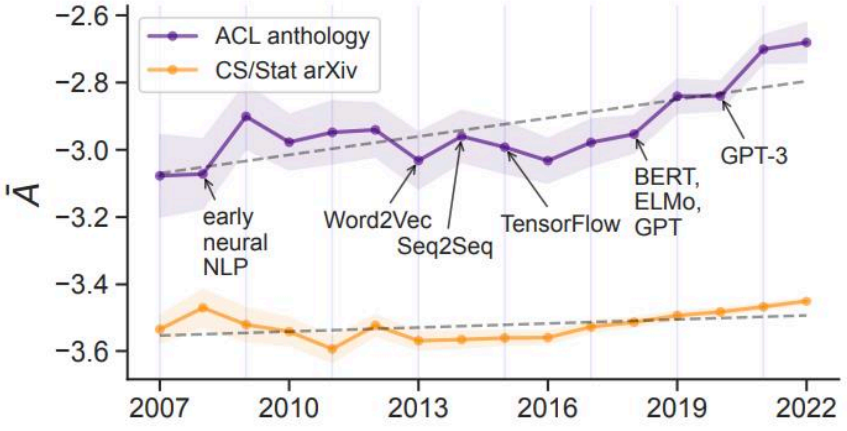


LM-related papers have higher scores than papers that do not mention LMs

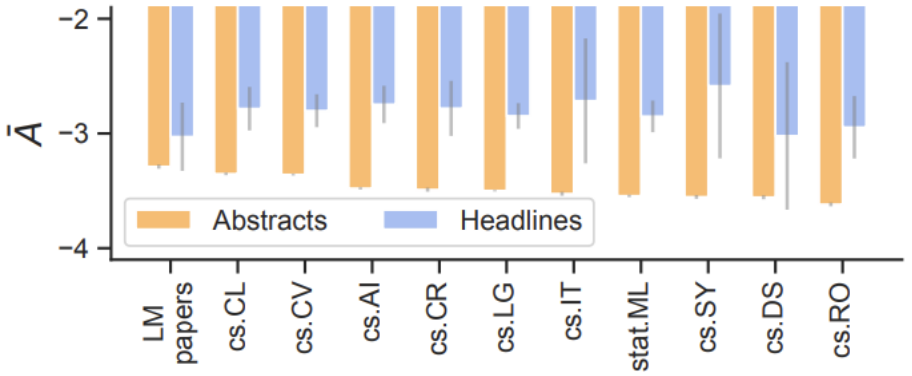


Within LM papers, LMs are much more anthropomorphized than other technical artifacts, but do not have as high of a score as human entities do

Additional Analyses



Anthropomorphism is increasing over time



News headlines anthropomorphize more than paper abstracts

Discussion

- Can you think of other scenarios where this use of MLMs might be useful?
  - Where else is measuring dehumanization or anthropomorphism useful?
  - What are other examples of metaphorical language?
  - What about other types of language?

- How might you improve the evaluation conducted in these projects?

- What are some of the limitations? How do they limit conclusions?

Acknowledgements

- Slide thanks to Daniel Khashabi: https://self-supervised.cs.jhu.edu/sp2024/