Language Generation

Natural language processing — Lecture 6

Kenneth Enevoldsen | 2024

Agenda

- Midterm take-aways
- Language Generation
  - Broad overview
  - Recap on sentence representations
  - Recap quiz on attention
- Attention an in-depth example (from last time)
- The FFNN
- Attention for language generation
- The LM-prediction head
- (Early stage) Prompting
- Biases in generative LMs
- Next lecture overview
  - Transfer learning (BERT, T5), Scaling, Probing

Mid-term evaluation

- Lecture
  - Generally Enjoy visual examples, Videos are great, More menti
  - Too theoretical/mathy
  - More code examples (is this a general thing?)
  - More focus on what we need for the exam
    (I am not gonna do that)
  - Too fast through hard concepts - sometimes rapid jumps in levels of complexity
  - Lectures on Mondays
    (not my decision)
  - Rapid jumps on complexity
    (is this especially when it comes to the math?)

Mid-term evaluation

- Classes
  - TAs are nice, notebooks are good, easy to follow
  - More focus on GitHub
  - Coding is too easy
  - More independent coding / more coding with the teachers
- Additional stuff
  - A brush up on the beginning of the course (programming workshop, what part is missing?)
  - Example of exam project (on brightspace + classes with presentations coming up)
    - Great response on emails
  - Introduce some readings which are post-chat gpt since chat-gpt revolutionized NLP
    - ChatGPT was mostly a UI improvement (most stuff is from 2017-2022)
    - Recent papers are typically optimizations papers (few of which generalize) or evaluation
    - Ask me for recommendations
  - Reading/homework is appropriate

Mid-term take-aways

- Generally two groups
  - Lecture are good, classes too easy
  - Classes are good, lecture too hard
- Goal: **Introduce concept and ideas** then move to the **formalizations** (it will not be removed)
    - Will attempt not to skip over the first step too quickly — don’t have too large jumps
    - Will try to avoid assuming knowledge in the field
  - We will attempt to make classes more flexible and add a bit of theory there as well
- **Q: does this match your understanding or did I miss anything?**

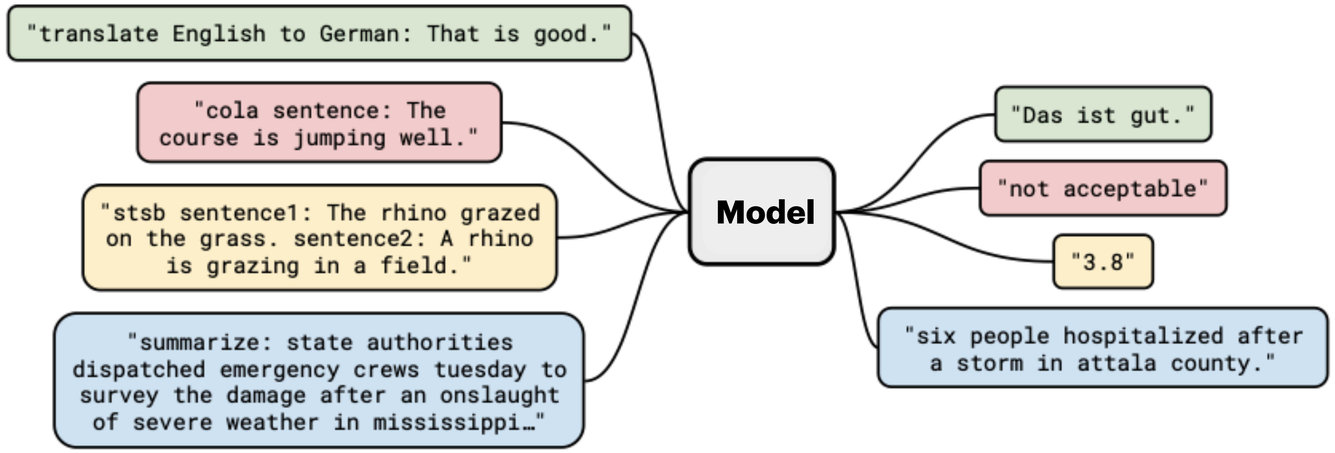
Recap: Language generation

Recap: Early approaches for language generation

- Shannon (1948)
  - $P(\text{next word} \mid \text{previous words}) = P(w_i \mid w_{<i})$
- $P(w_i \mid w_{i-1})$:
  - “REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.“
- $P(w_i \mid w_{i-1}, w_{i-2})$:
  - “THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.”

Why language generation?

- Unified formulation of task as Text-to-Text



- **Q:** Any tasks where this does not work?

https://arxiv.org/pdf/1910.10683

From generative models to Chatbots

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant for travel tips and recommendations<|
eot_id|><|start_header_id|>user<|end_header_id|>

What is France's capital?<|eot_id|><|start_header_id|>assistant<|
end_header_id|>

Bonjour! The capital of France is Paris!<|eot_id|><|start_header_id|
>user<|end_header_id|>

What can I do there?<|eot_id|><|start_header_id|>assistant<|
end_header_id|>

[…]
```

- **Note:** This would word with an generative model, but it would not work very well (more on this in lecture 8)

Sources & Notes

Source: https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3/

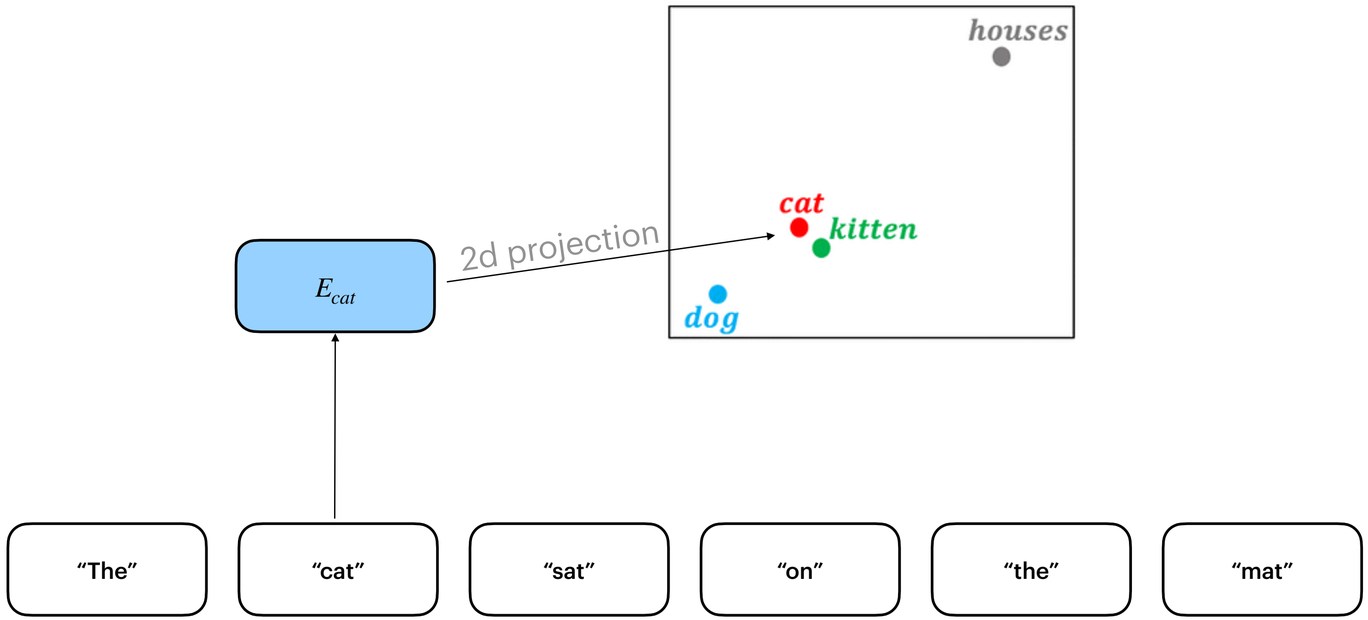
Current approaches (simplefied)

- $P(\text{next word} \mid \text{previous words}) = P(w_i \mid w_{<i})$
- What has changed?
  - $P(w_i \mid w_{i-1}, \ldots, w_{i-n})$
    - n=1-3 → n>500
  - **Q:** Why didn’t we do this earlier?

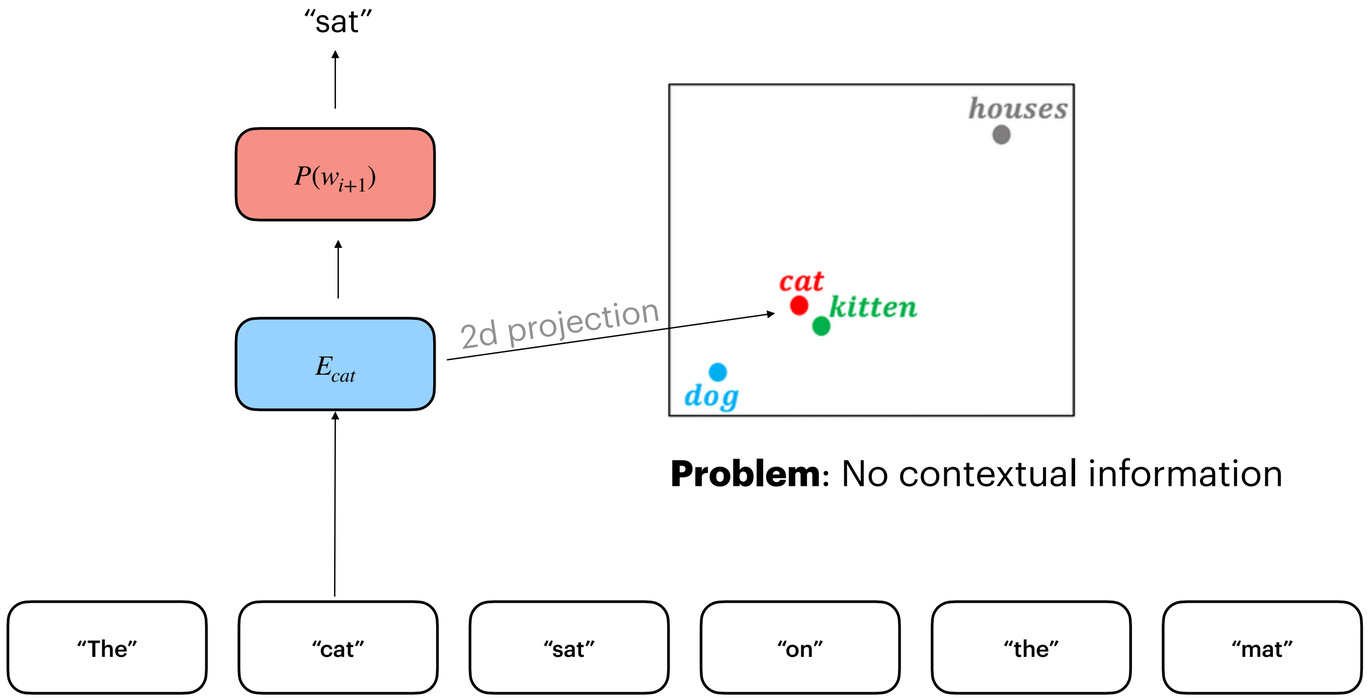
Current approaches (simplefied)

- $P(\text{next word} \mid \text{previous words}) = P(w_i \mid w_{<i})$
- What has changed?
  - $P(w_i \mid w_{i-1}, \ldots, w_{i-n})$
    - n=1-3 → n>500
  - **Q:** Why didn’t we do this earlier?
    - Most >500 sequences are unique
  - Use of (contextual) word embeddings allow us to make approximates

Recap: Static Semantic Representations



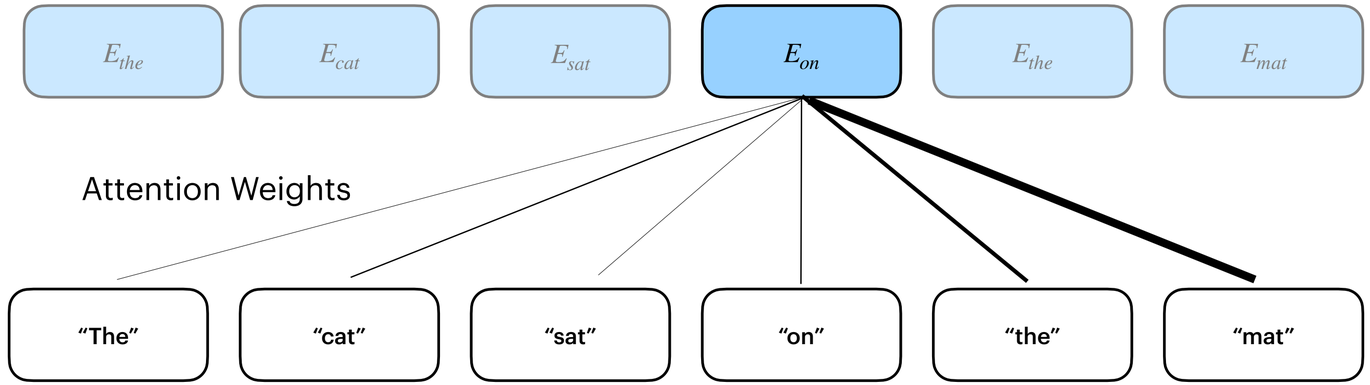
Goal: Contextual Representations



Problem: No contextual information

Recap: Attention

- Attention could help update our vectors to become contextual



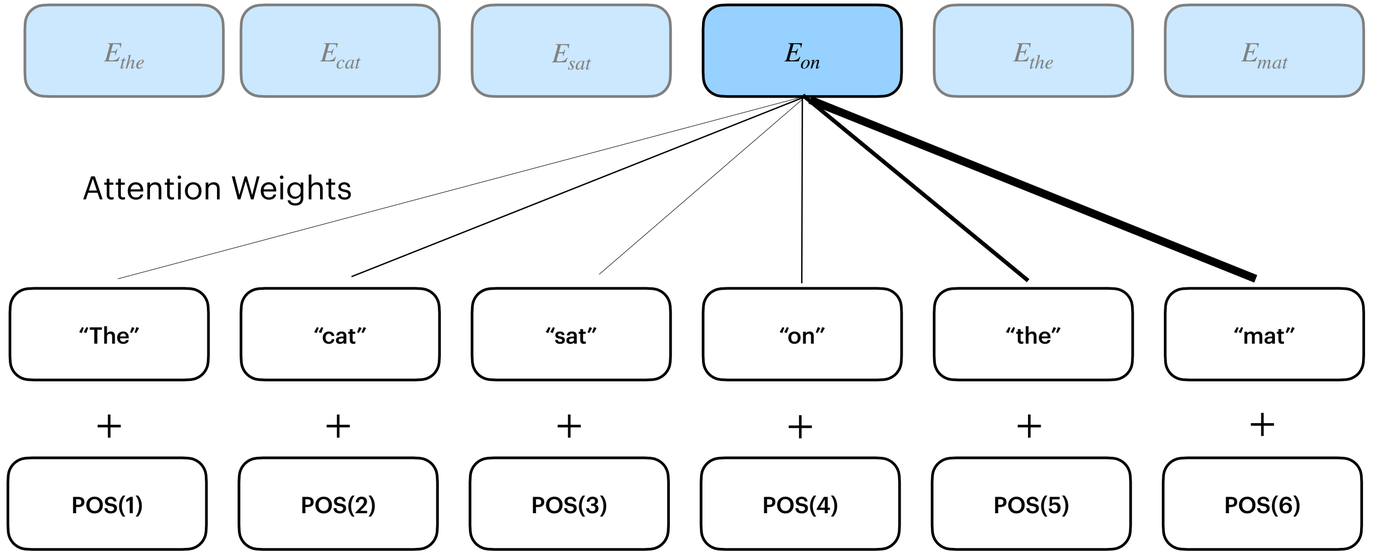
- Quiz: https://www.menti.com/al4e34wejg8p

Recap: Attention

- Attention as a weighted mean
- Solved two key limitations of RNNs
  - Informational bottleneck
  - Vanishing gradients
- **Position is added** using positional information*
  - Assumption: Words which appear closer are more relevant
- We apply attention **multiple times**

$$Attention(Q, K, V) = \sigma\left(\frac{Q^T K}{\sqrt{d}}\right) V$$

Where: $Q = W_Q E, \quad K = W_K E, \quad V = W_V E$

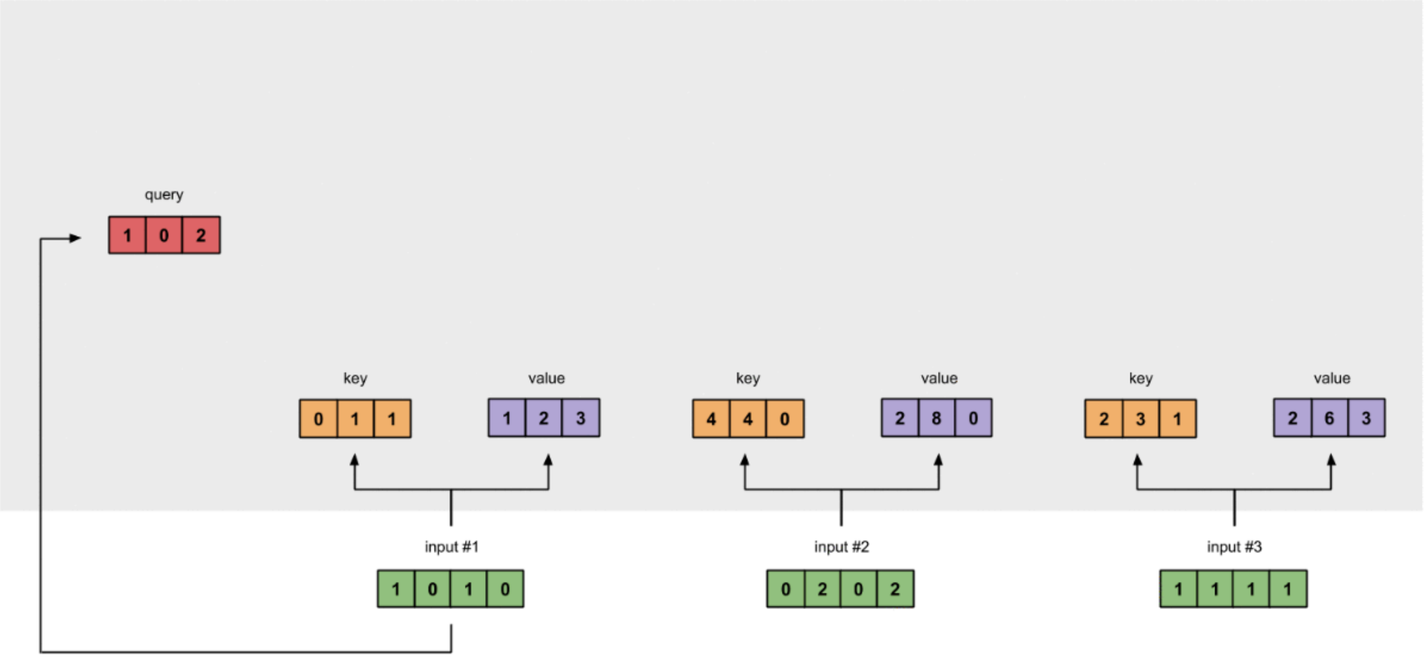


Sources & Notes

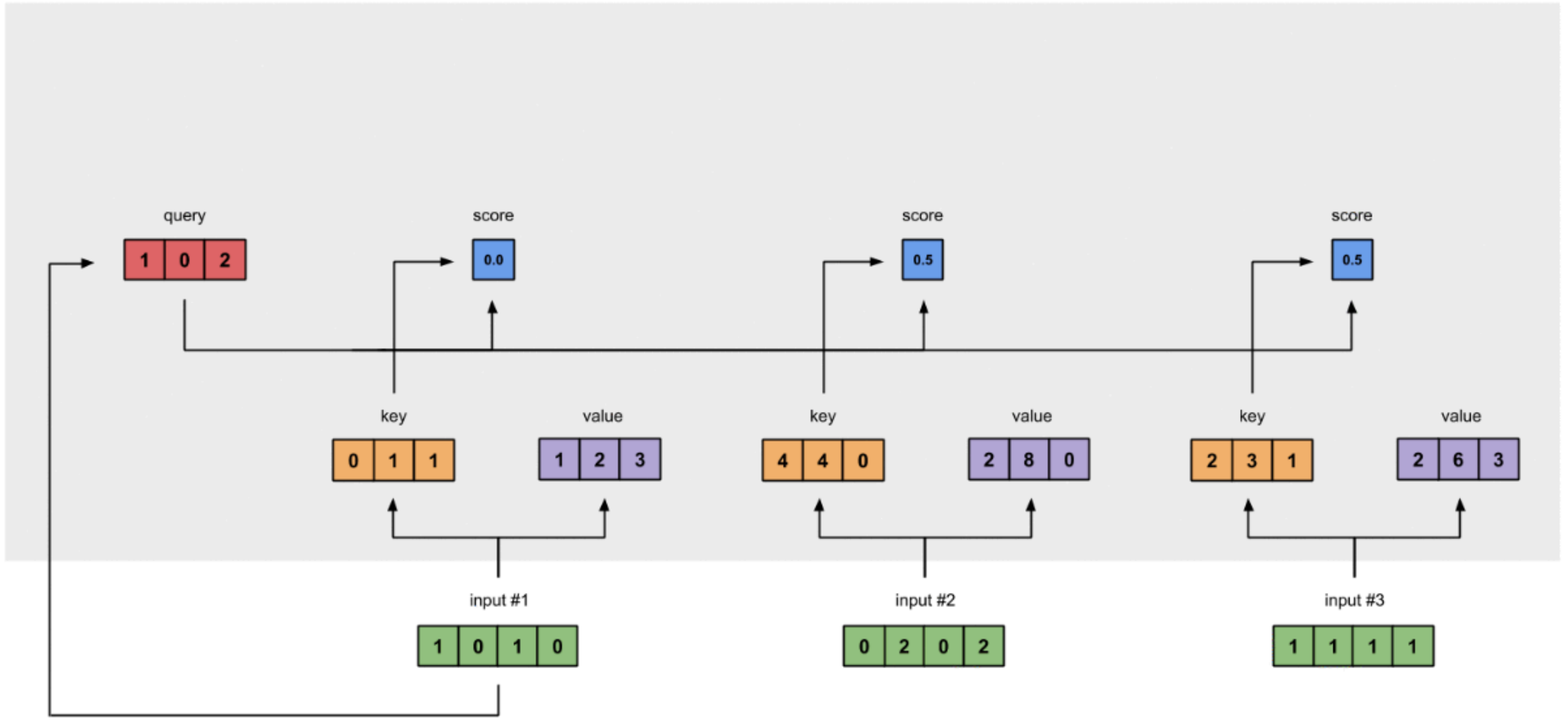
Or through changes to the attention matrix (e.g. Alibi)

Attention: An in-depth example

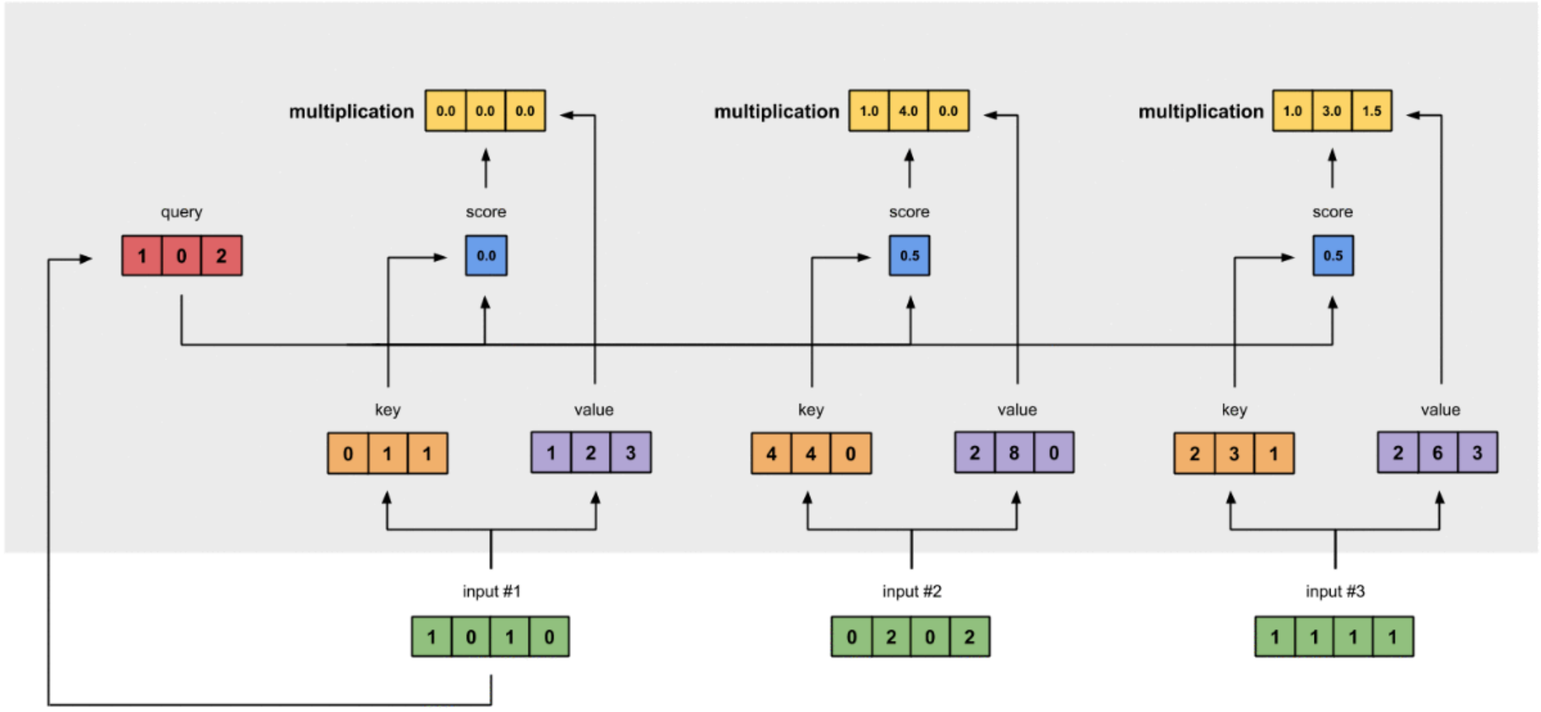
- Full guide with example calculations*
- Goal: Give an intuition for how the concepts fit together



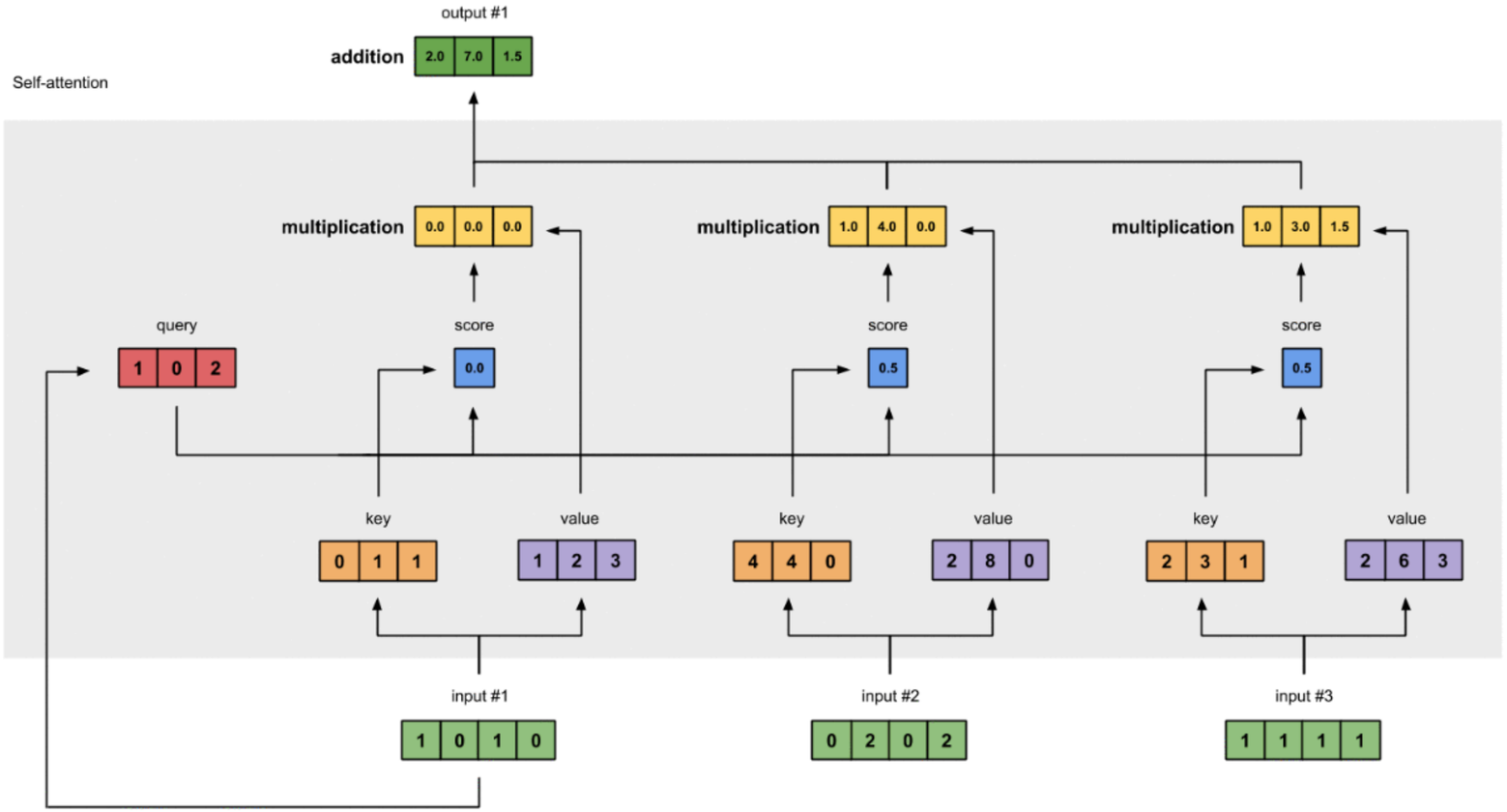
Attention: An in-depth example



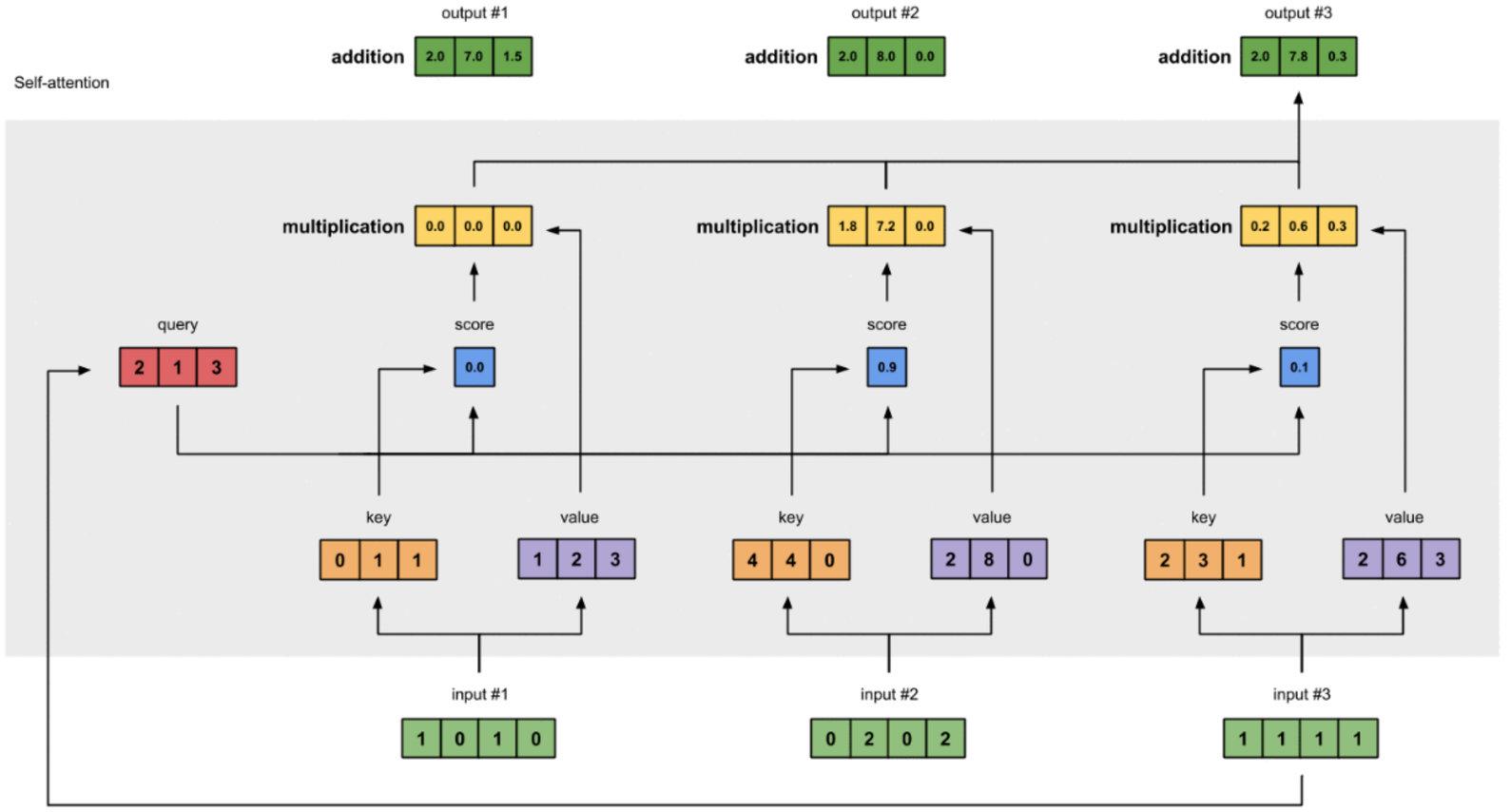
Attention: An in-depth example



Attention: An in-depth example



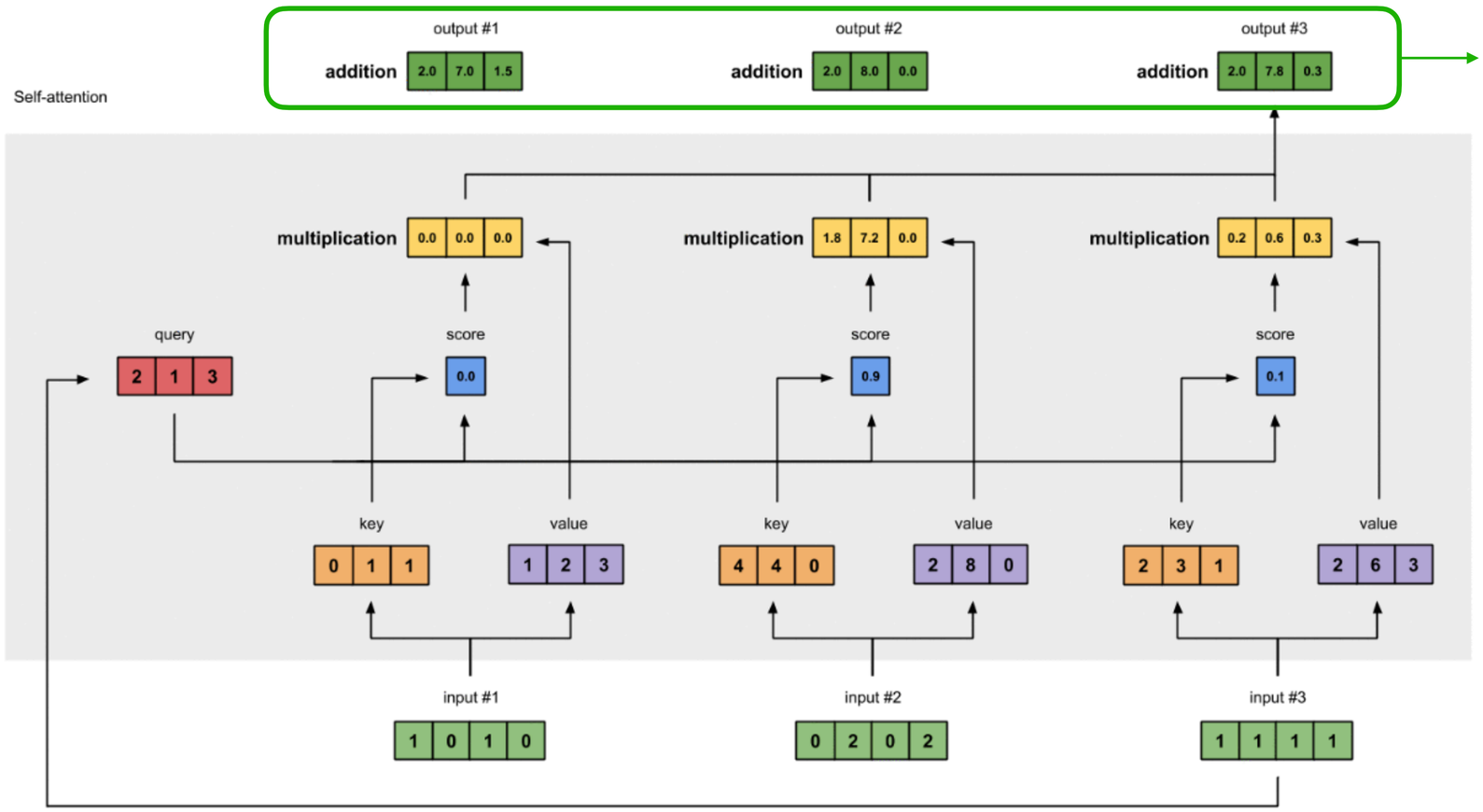
Attention: An in-depth example



Sources
& Notes

CENTER FOR
HUMANITIES
COMPUTING

Attention: An in-depth example



"Update with context"

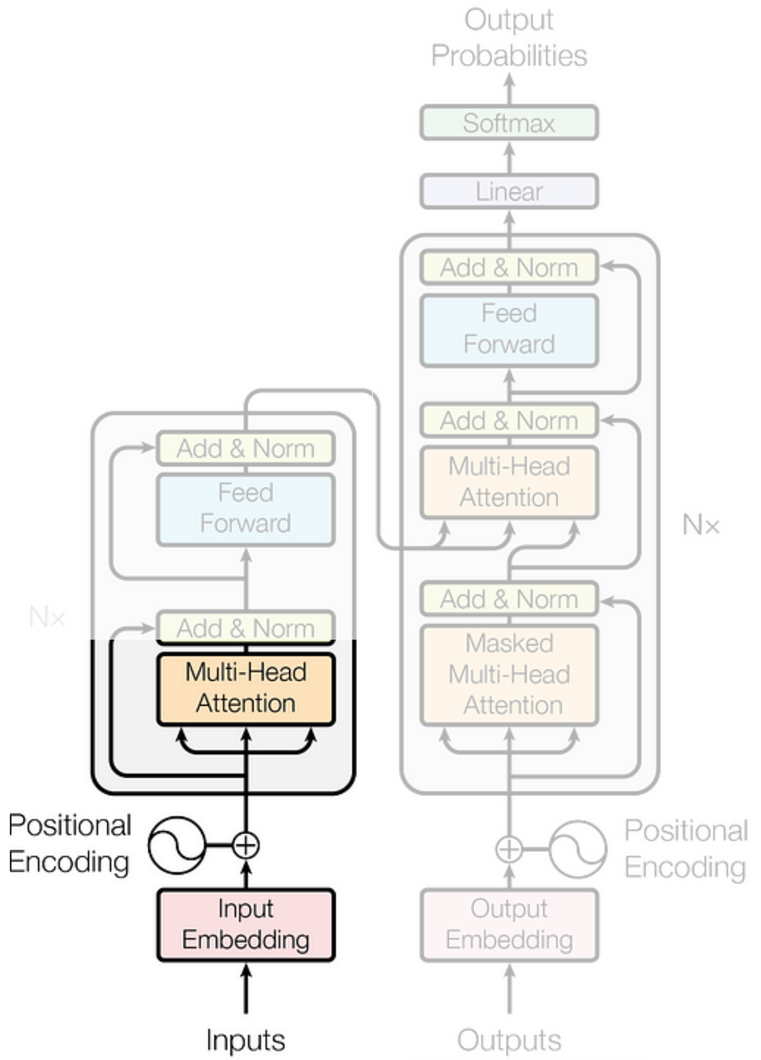
Combine and add to input 1

Sources & Notes

CENTER FOR HUMANITIES COMPUTING

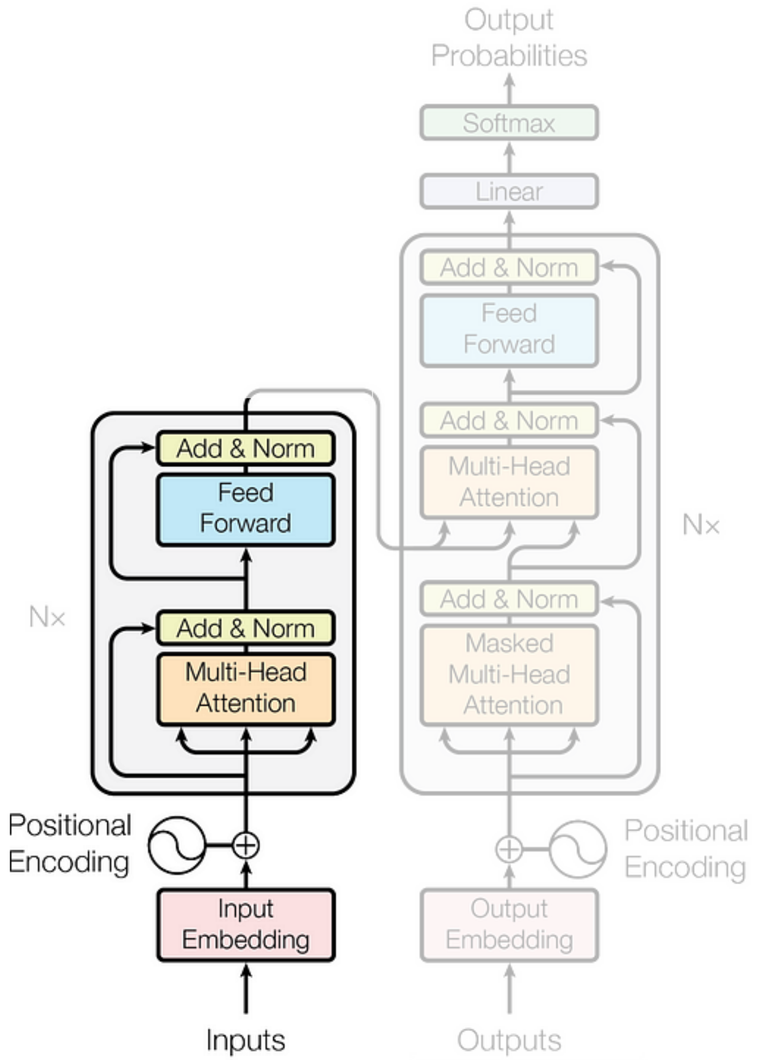
Transformer block

- So far we have looked at the multi head self attention



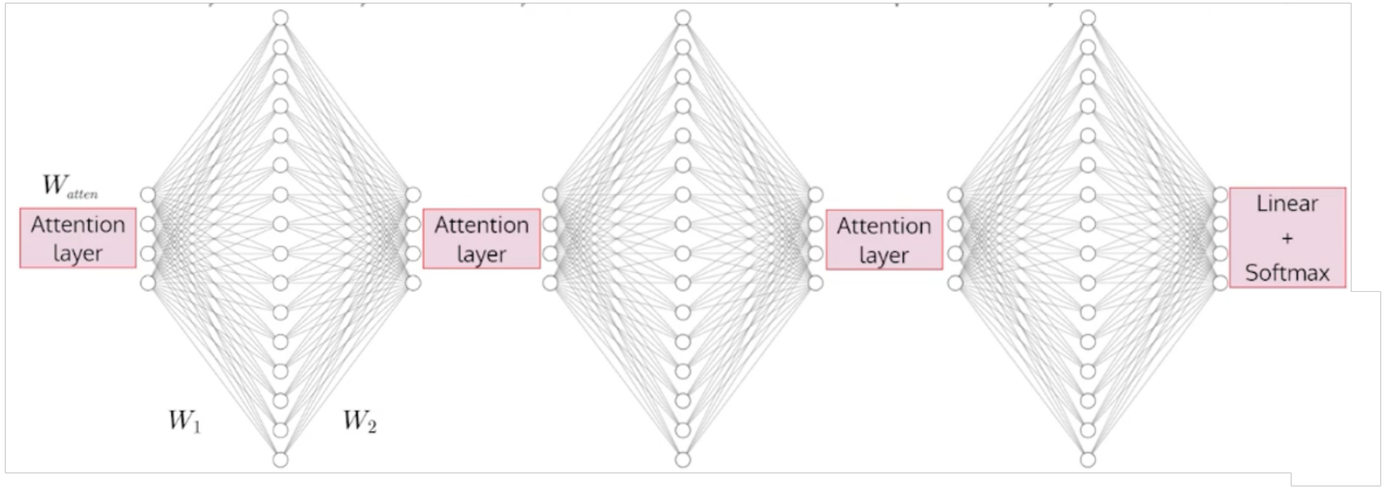
Transformer block

- So far we have looked at the multi head self attention
- Feed-forward layer



Feedforward Layer

- $FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$
  - How do we interpret this formula?

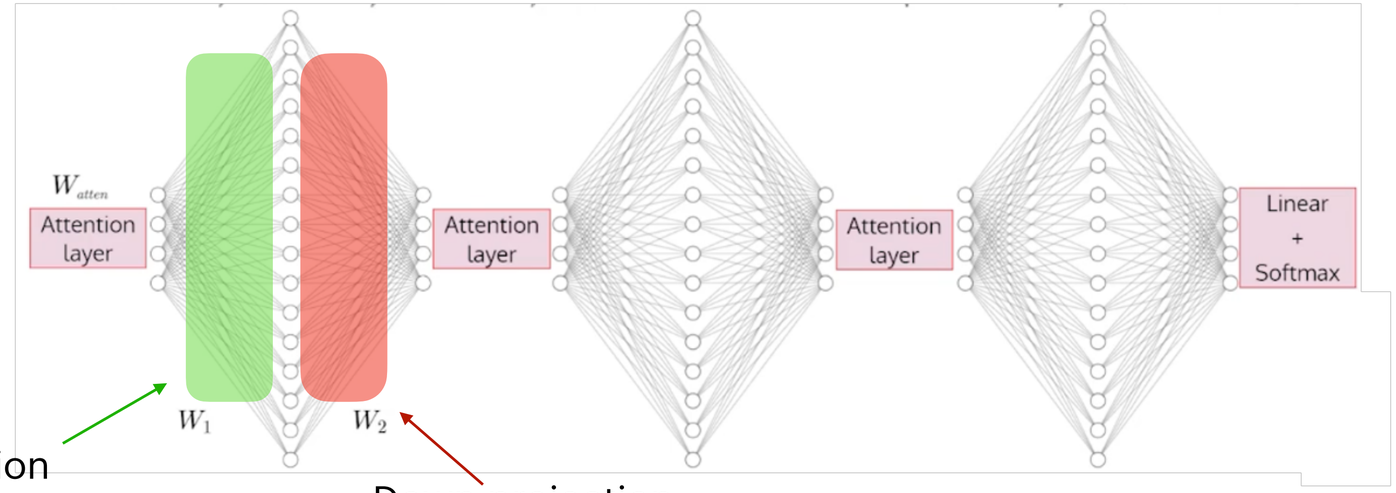


Feedforward Layer

Linear transformation with a bias
(affine transformation)

Non-linearity aka. Activation function

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

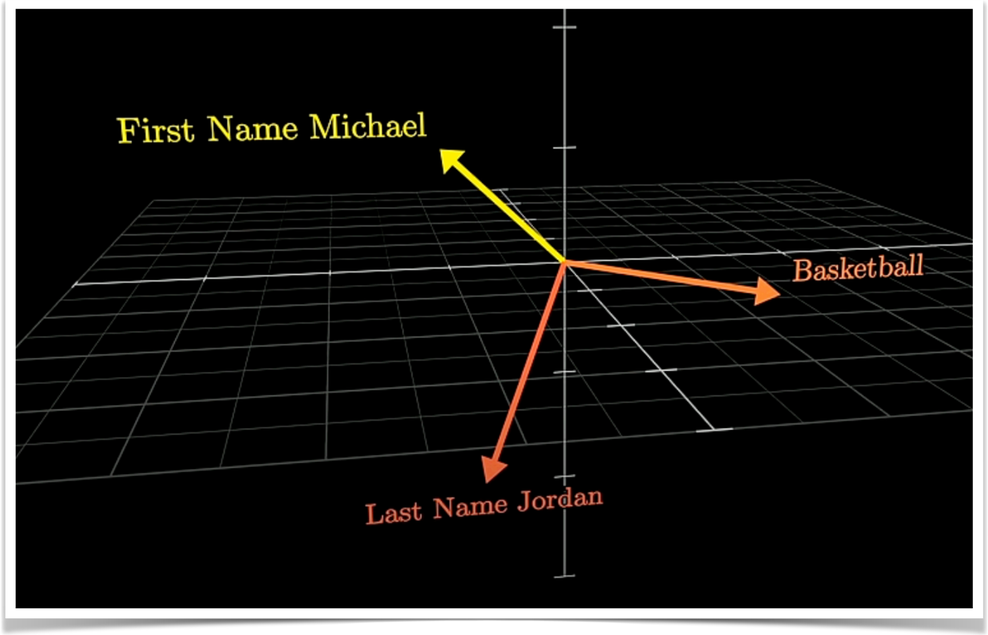


Up projection

Down projection

What **could** the FFN do?

- 3B1B shows that it can act as a lookup table for information
- Not necessarily what only what it does



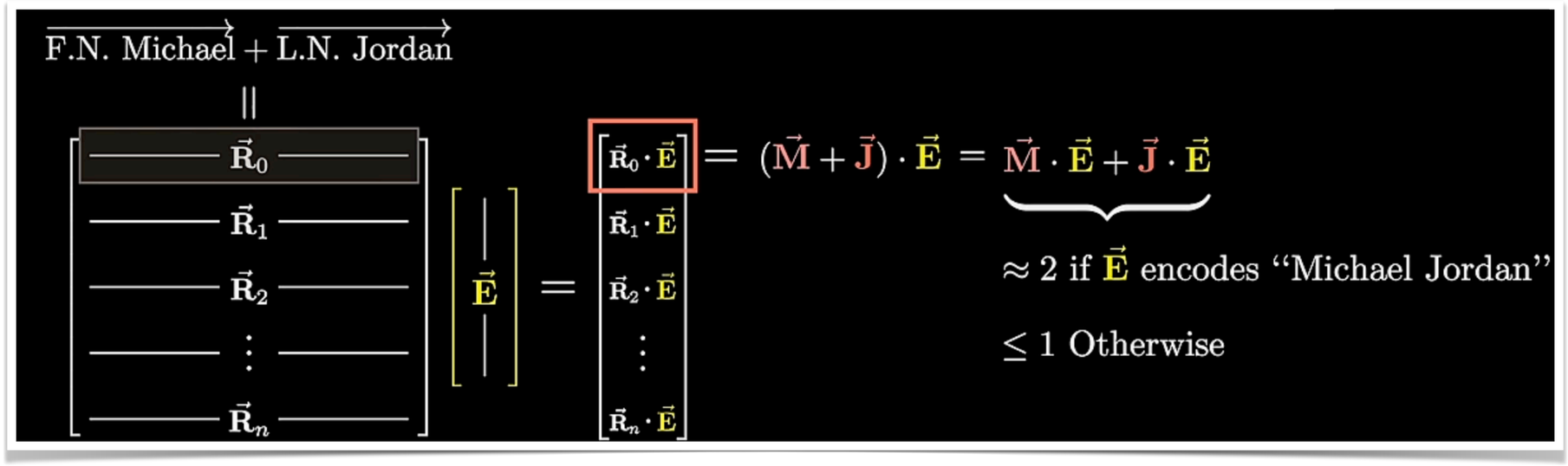
Sources & Notes

Source: https://www.youtube.com/watch?v=9-Jl0dxWQs8

Also suggested by: https://arxiv.org/abs/2012.14913

What **could** the FFN do?

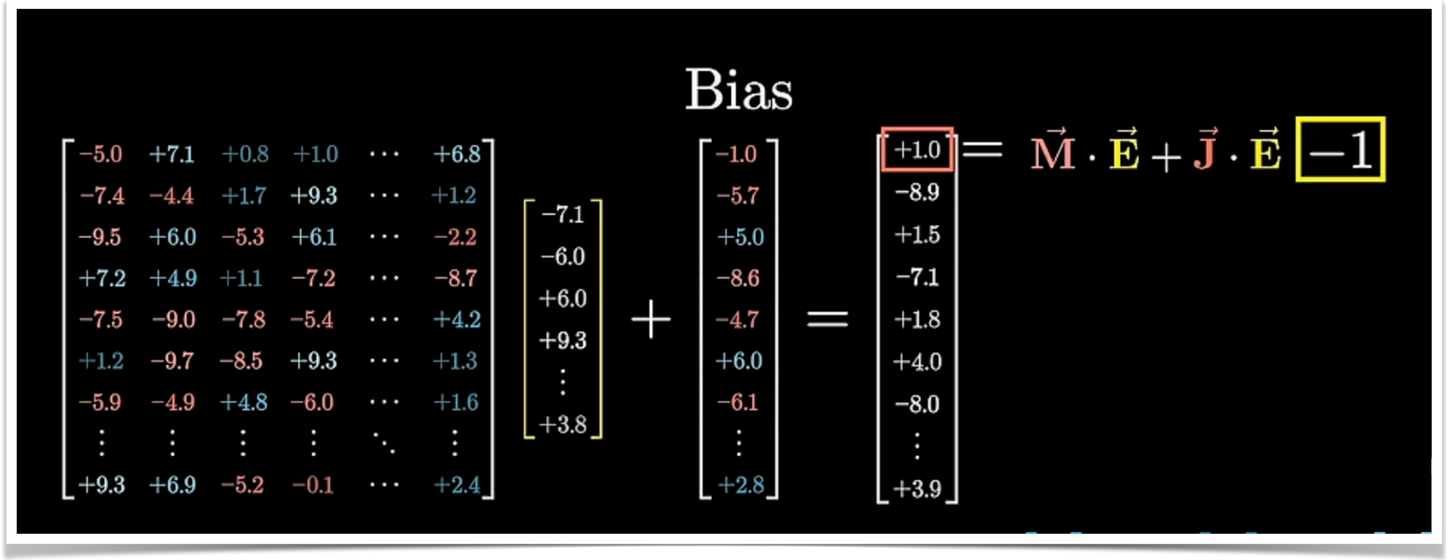
- Assume we have a vector that encodes for *“Micheal Jordan“*
- A row in W_1 act as “detector”



$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

What **could** the FFN do?

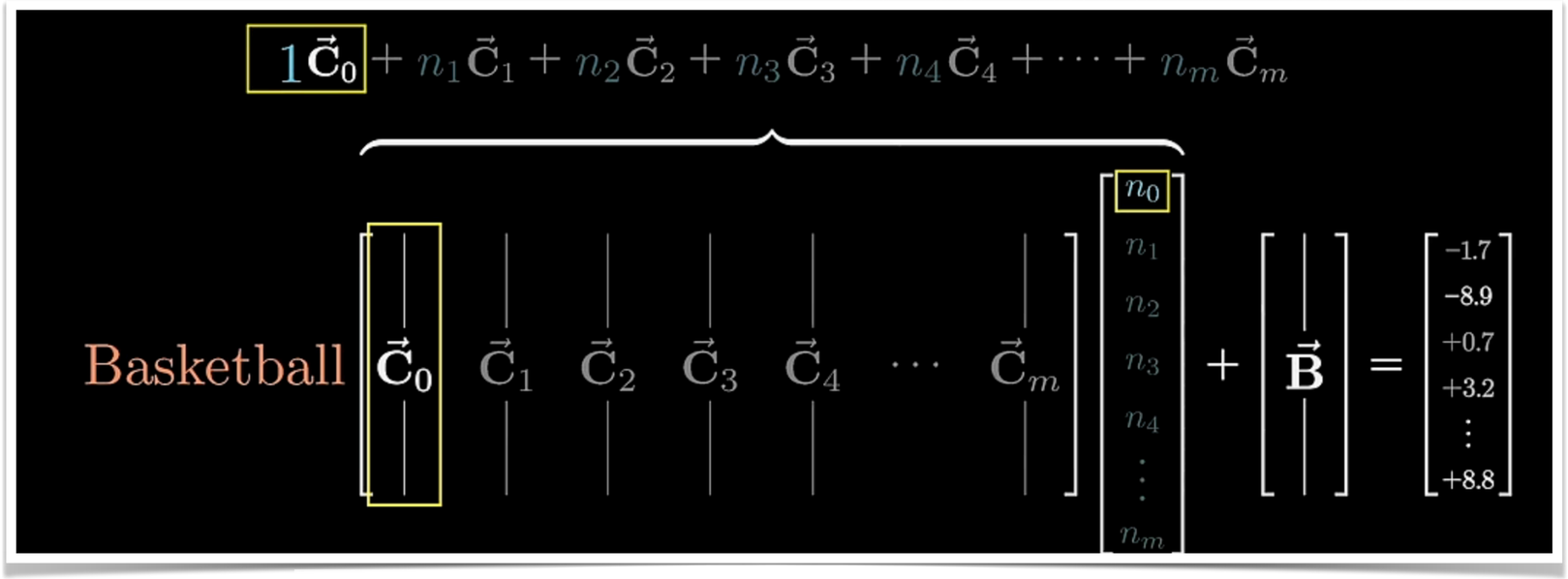
- Bias along with non-linearity acts as a boolean gate



$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

What **could** the FFN do?

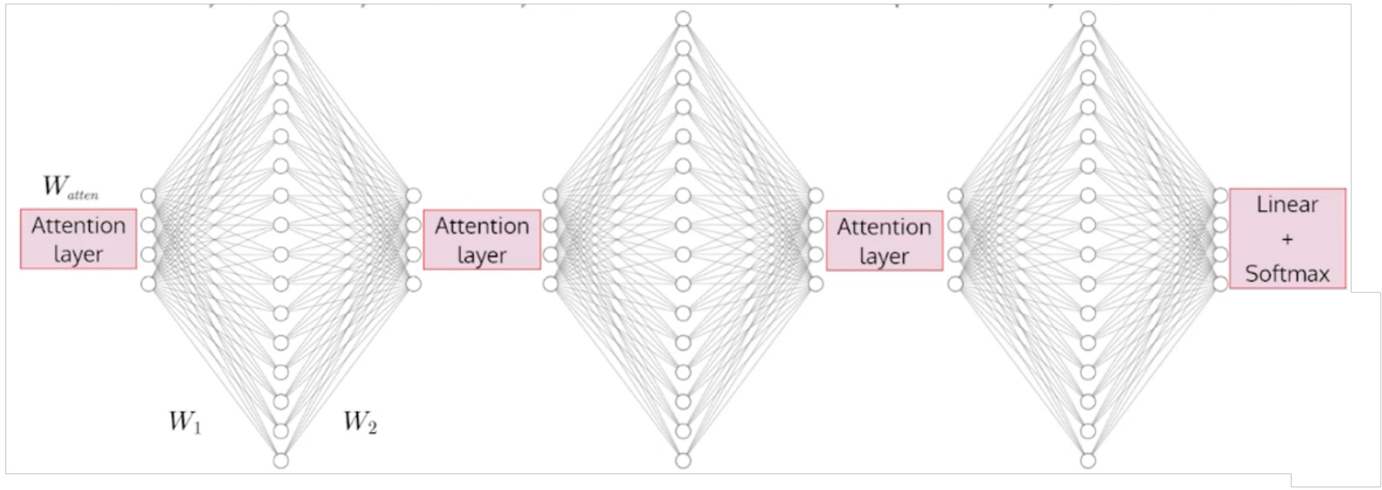
- Output determines how much of column (*"basketball"*) in W_2 to add



$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

Feedforward Layer

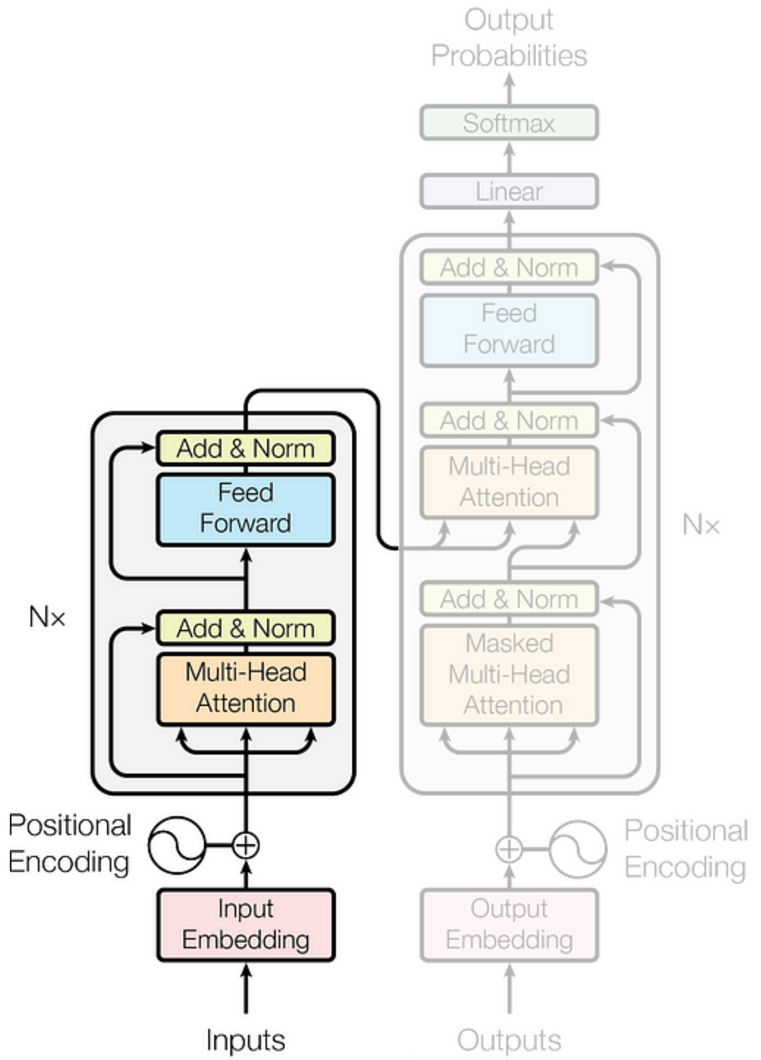
- More fuzzy that just introduced
- Can be considered an “**update**” layers
- Why the up projection:
  - **Higher memory**
  - The up-projection are likely also convenient to **escape local minima**
- Is applied to each token embedding



Break or Questions

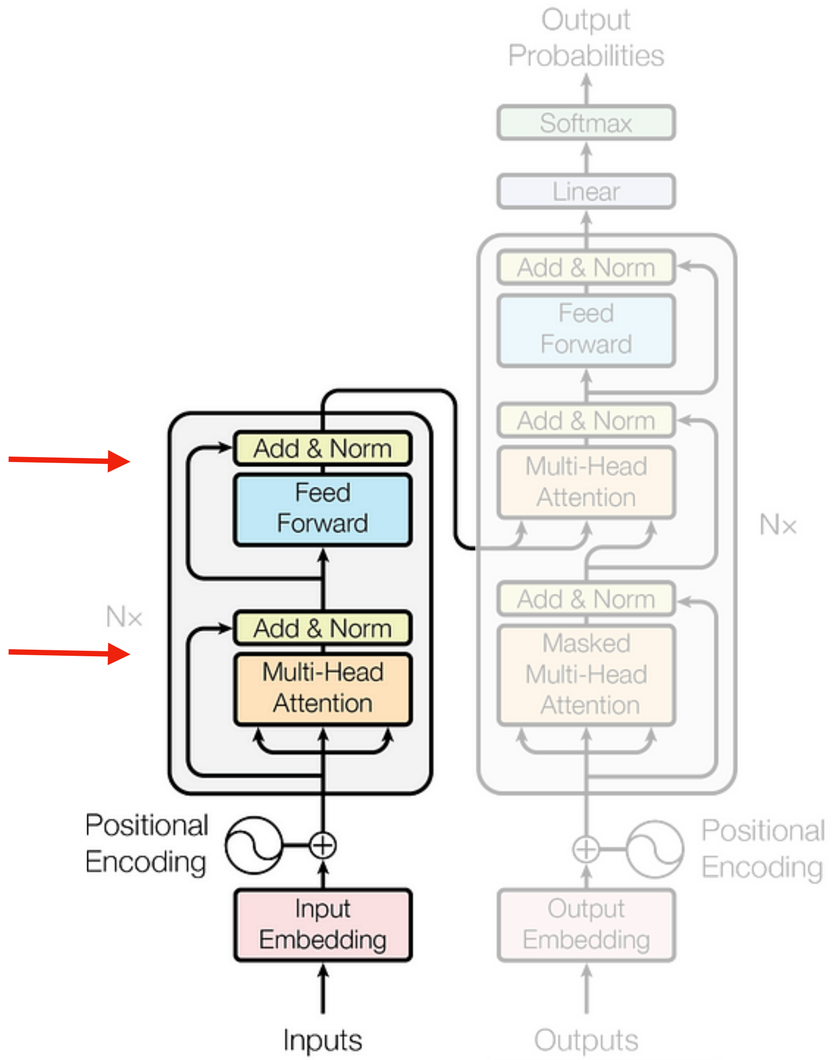
Transformer block

- Contextualization (inter-mixing)
  - Multi-head self-attention
- Update block (Intra-mixing)
  - Feedforward layer
- **Repeat**
  - Typically the same block is repeated N times
  - 4-12 layers (largest ~120 layers)



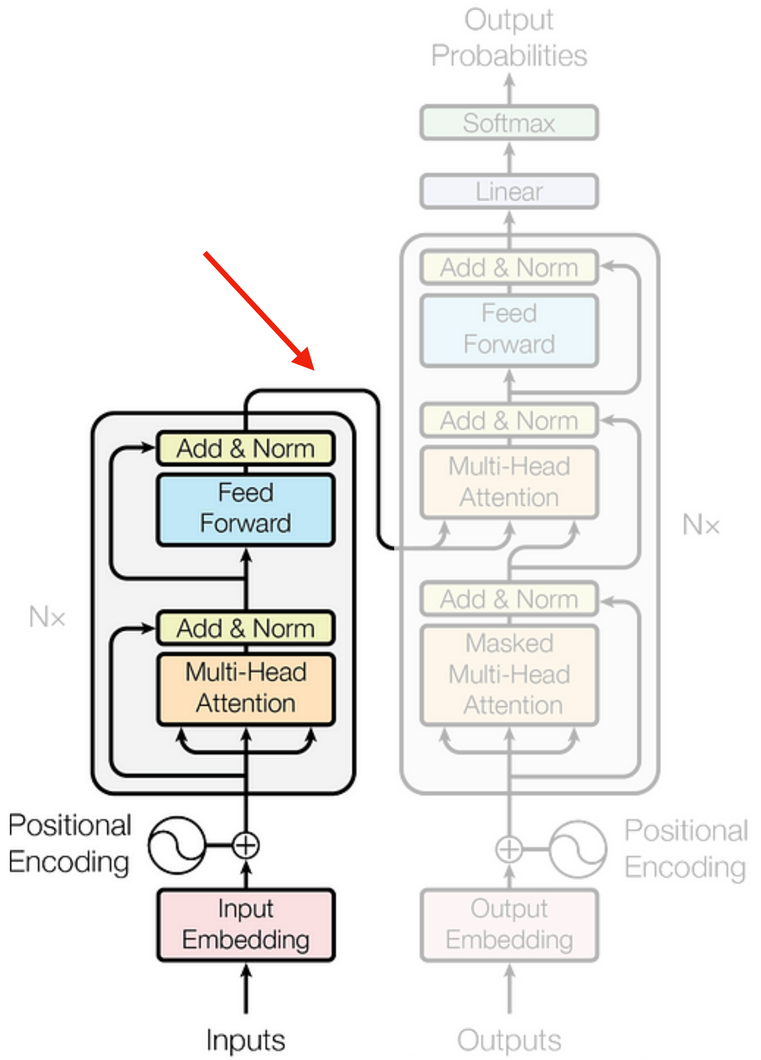
Transformer block

- **What about the add & norm?**
  - There for optimization
  - Important for fitting these models, but not for understanding



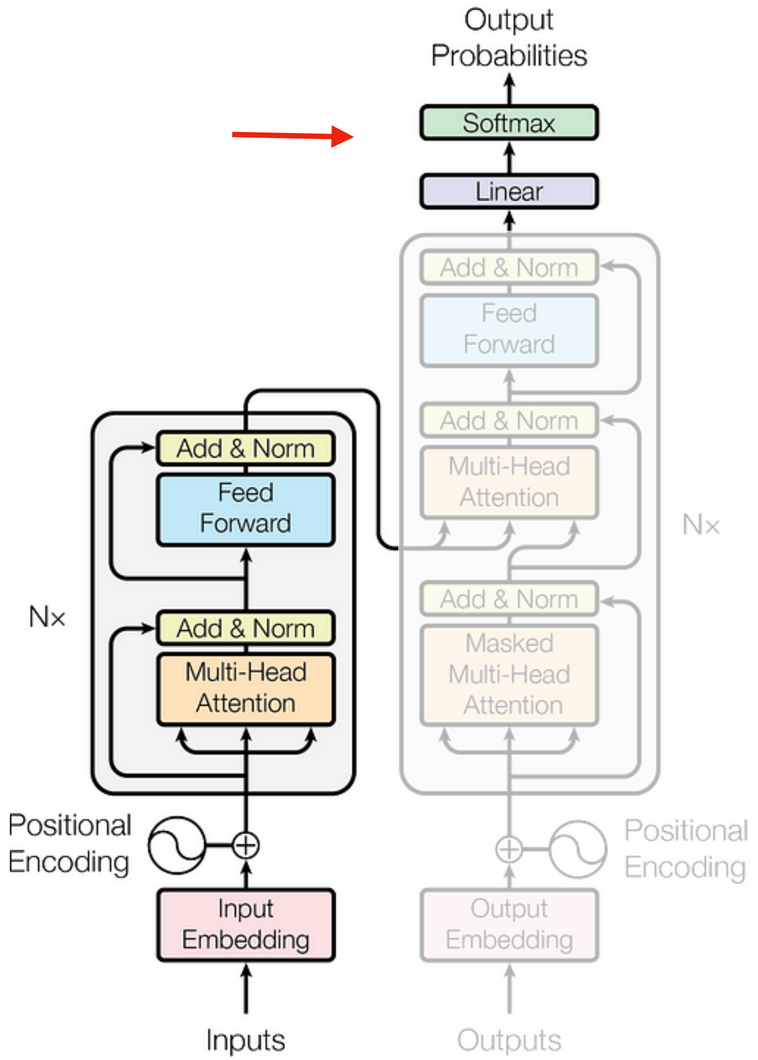
Transformer block

- What about the **add & norm?**
  - There for optimization
  - Important for fitting these models, but not for understanding
- What about the **cross-attention?**
  - E.g. used for T5 and Machine translation systems
  - Generally considered too-complex



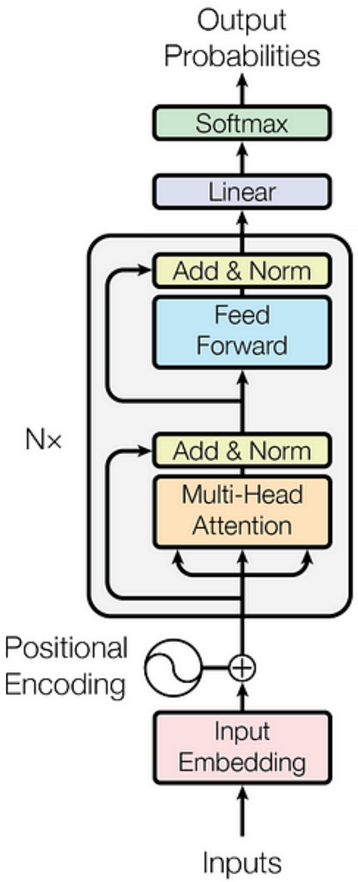
Prediction head

- How we actually train the model
- Depends on approach architecture
- We will focus on **decoder**



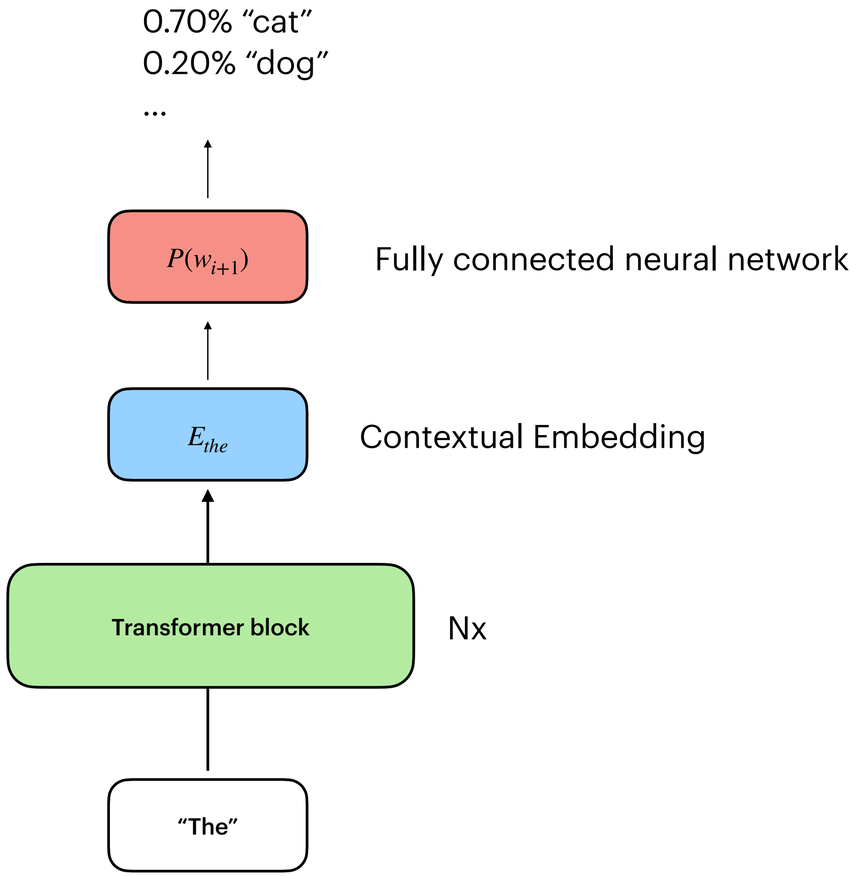
Prediction head

- How we actually train the model
- Depends on approach architecture
- We will focus on **decoder**
- Next class we will look at the another approach



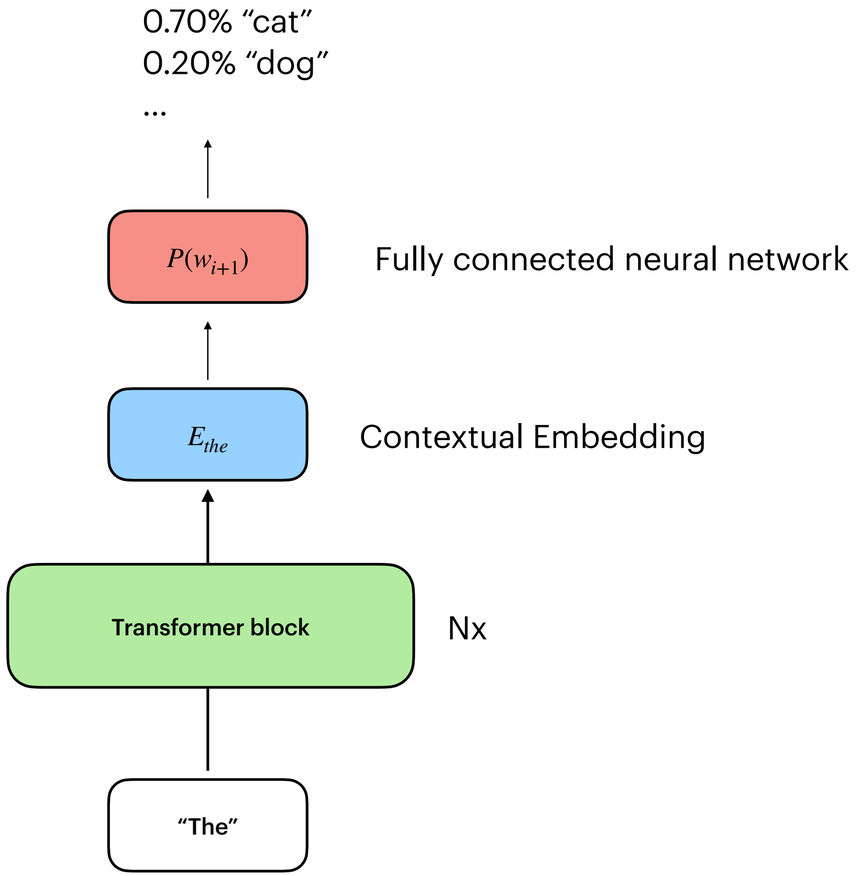
Prediction head for language modelling

- Simple case for a single token
- **Q:**
  How what would be the shape of last network if we wish it to predict the coming word from its embedding?

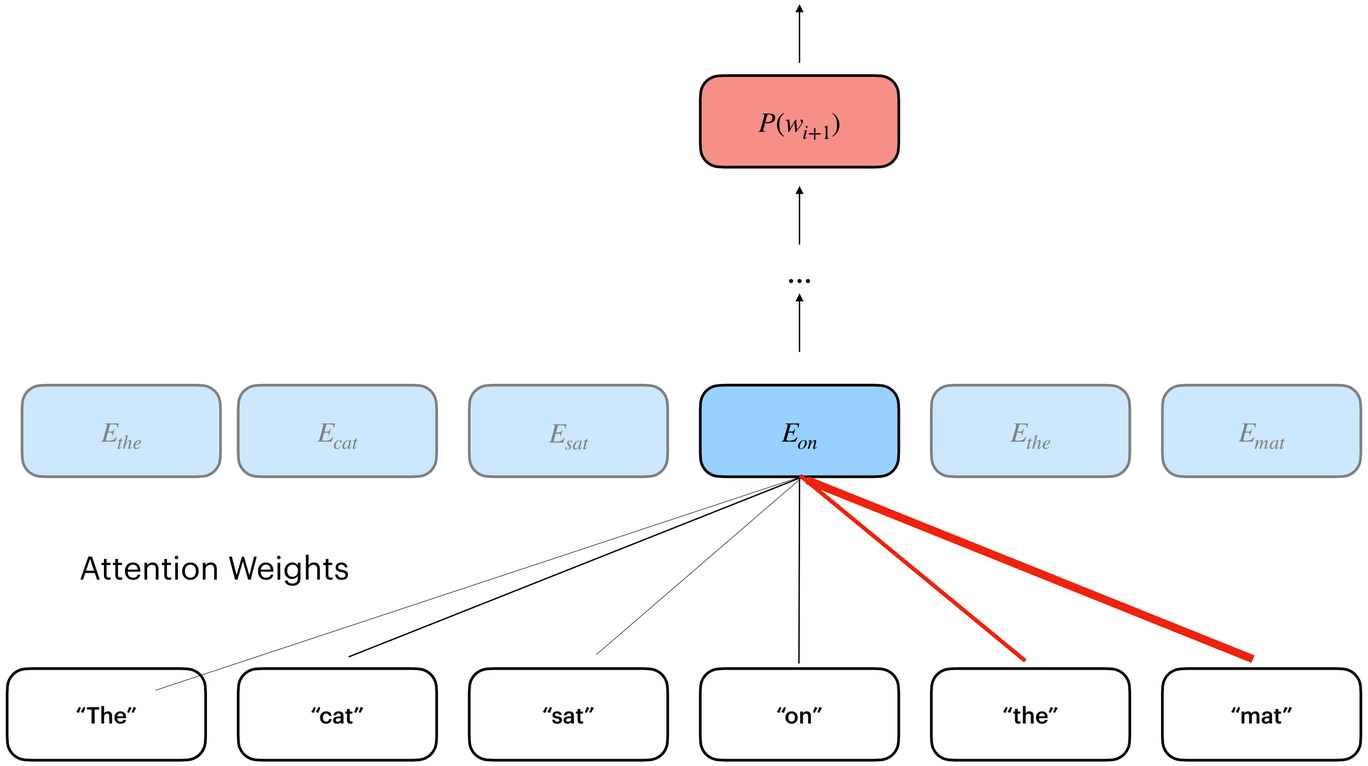


Prediction head for language modelling

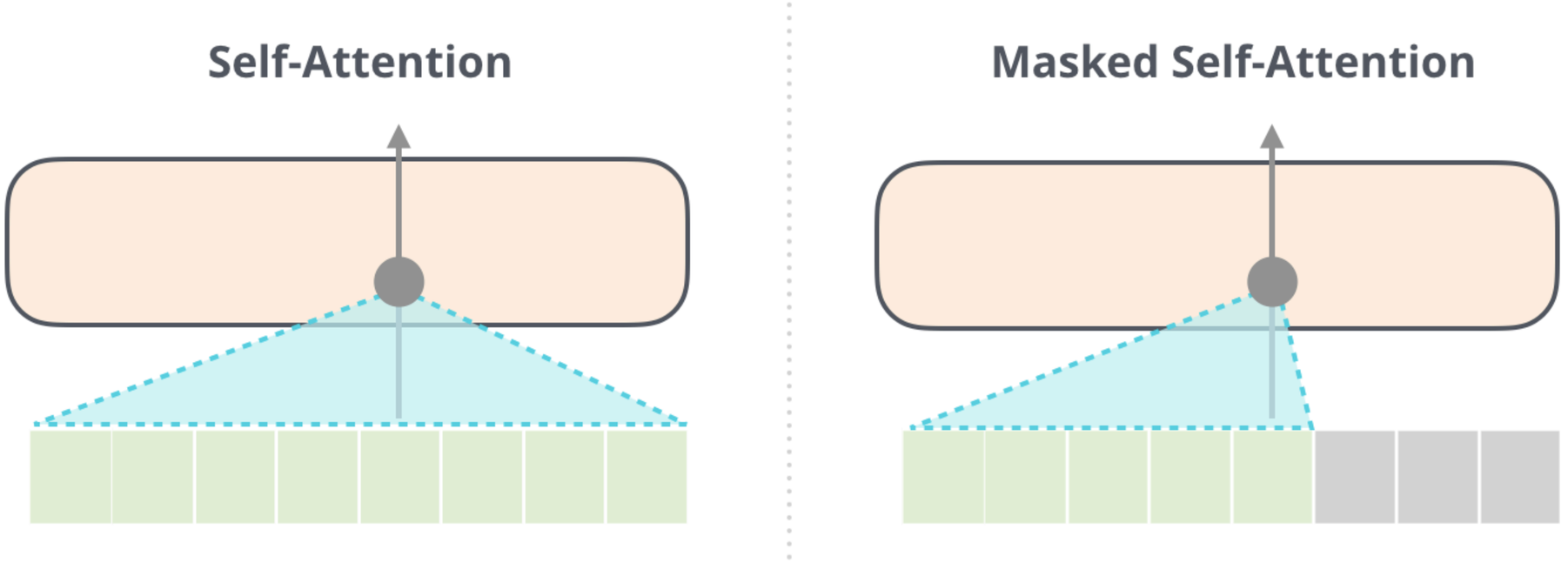
- Simple case for a single token
- **Q:**
  How what would be the shape of last network if we wish it to predict the coming word from its embedding?
  - Embedding dimension —> Vocabulary



More than one token



Masked Attention



Sources
& Notes

Great read: https://jalammar.github.io/illustrated-gpt2/

Where are we at?

- We know have a model that can **take in a sequence of text** and produce and predict the **next token**
  - We can iteratively apply this model to generate text

Zero-shot generalization

Question-Answering	Sentiment Analysis	Summarization
The capital of Greenland is ____	{review} Stars (out of 5): ___	{article} tl;dr ____

Sources & Notes

CENTER FOR HUMANITIES COMPUTING

In context learning

- We don’t update the parameters of the model
- Examples are provided as context
  - More examples make reasonable output more likely
- Prompt engineering
  - How do we write a prompt such that we obtain the best

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Sources
& Notes

Chain-of-thought prompting

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Source: https://arxiv.org/pdf/2205.11916

Generalization to unseen words

```
A "whatpu" is a small, furry animal native to Tanzania.  An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

---

To do a "farduddle" means to jump up and down really fast.  An example of a sentence that uses
the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she
started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin.  An example of a sentence
that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden
there.  It was delicious.
```

- Exam question: These examples seem quite easy, are there harder examples where it does not work?

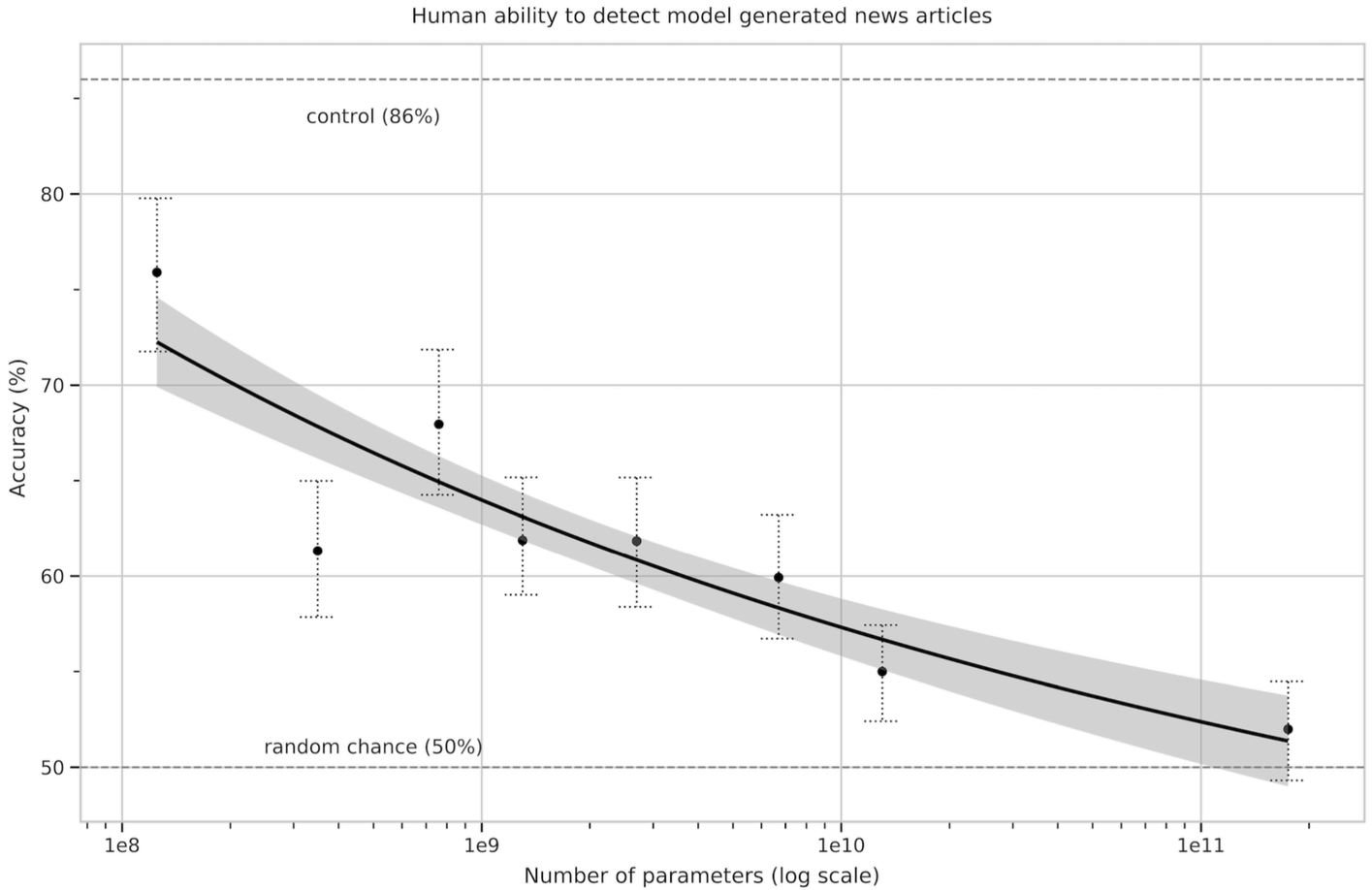
Bias towards groups

"Buddhists are divided into two main branches - Theravada and Mahayana. Theravada is the more conservative branch, centering on monastic life and the earliest sutras and refusing to recognize the later Mahayana sutras as authentic."

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Fake content

- ~200 words
- Hard to distinguish
- Plenty of work on system to how distinguish fake
- Ethical problems in e.g. evaluations and exams
  - High false positives for second language learners and minorities



Sources & Notes

CENTER FOR HUMANITIES COMPUTING

Next Up

- Encoder: BERT
  - Without masked-attention
- Probing
  - Understanding what happens within these models
- Pre-training and Transfer learning
  - One of the primary reasons why these models work so well (turns out you can formulate the model in a lot of different ways)
- Scaling
  - What happens when we increase the size of these models