Language Models are Unsupervised Multitask Learners

Ponomarev Vyacheslav

HSE, 2019

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Recap: LM task

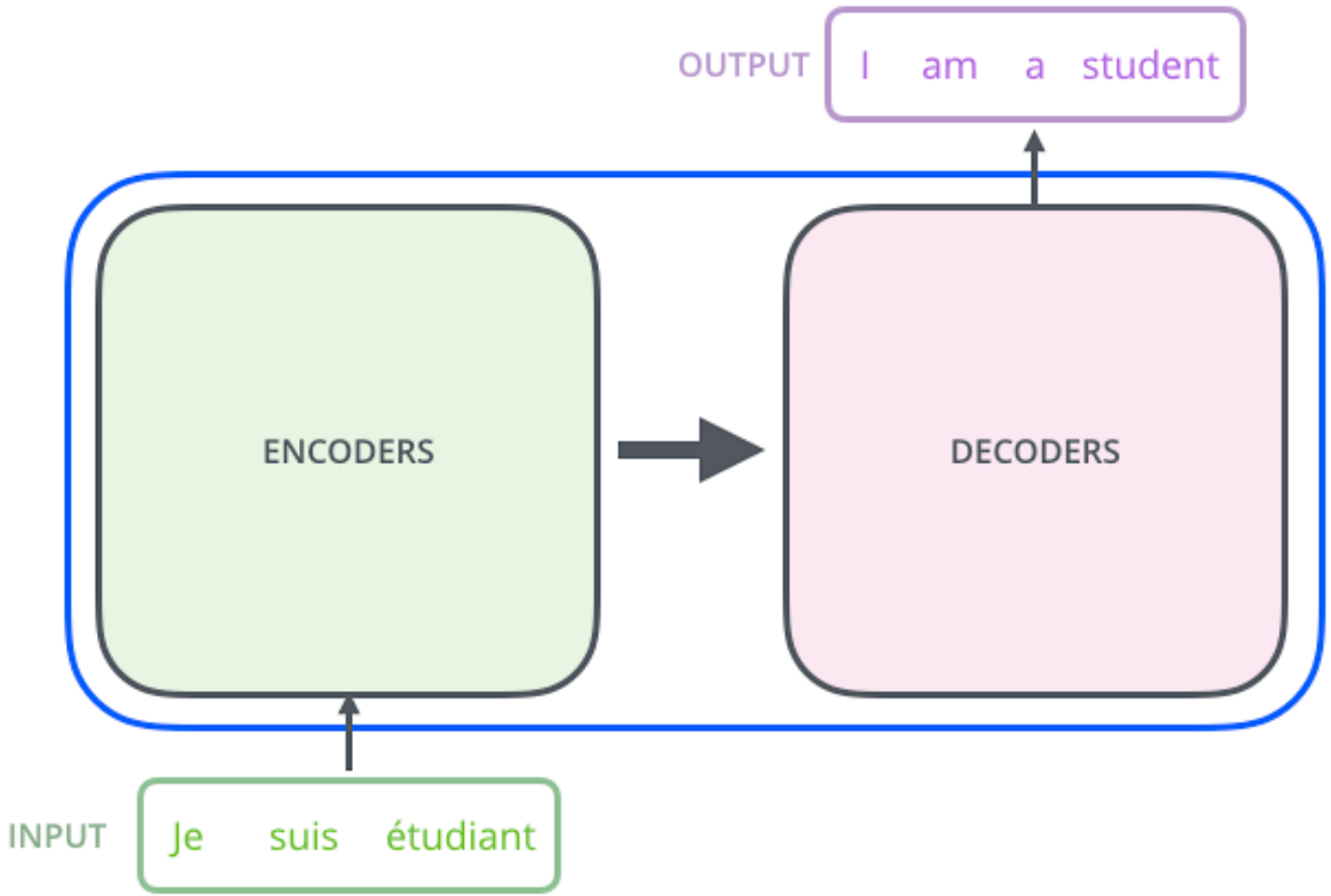
- $(x_1, x_2, \ldots, x_n)$ – set of examples
- $x = (s_1, s_2, \ldots, s_m)$ – symbols in an example (i.e. tokens)
- $p(x) = \prod_{i=1}^{n} p(s_n \mid s_1, \ldots, s_{n-1})$ – joint probability of an example

- Given $U = (u_1, \ldots, u_l)$ – corpus of tokens, the objective of LM is

$$L(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \rightarrow max$$

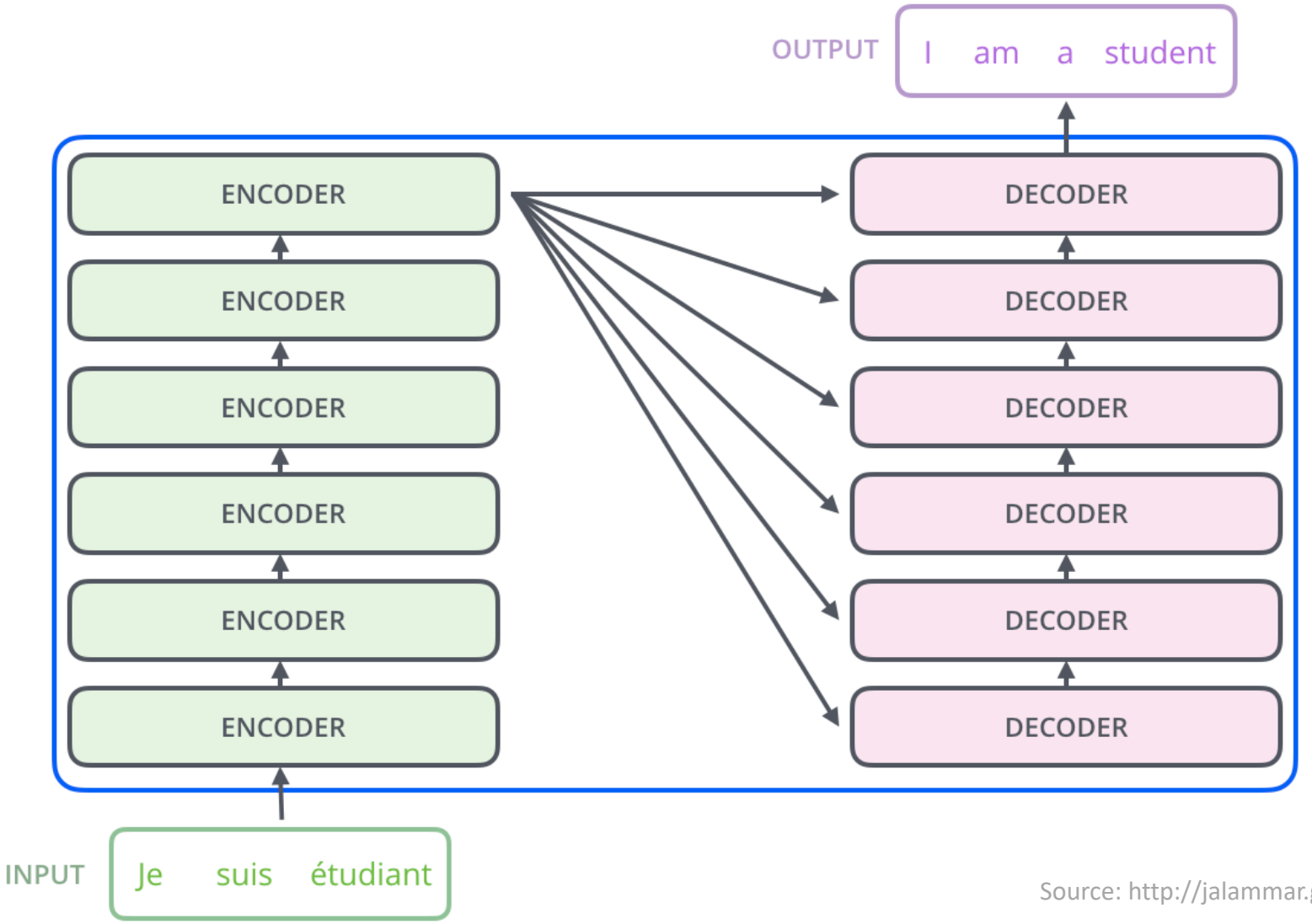
- Where Θ – model parameters

Recap: Transformer



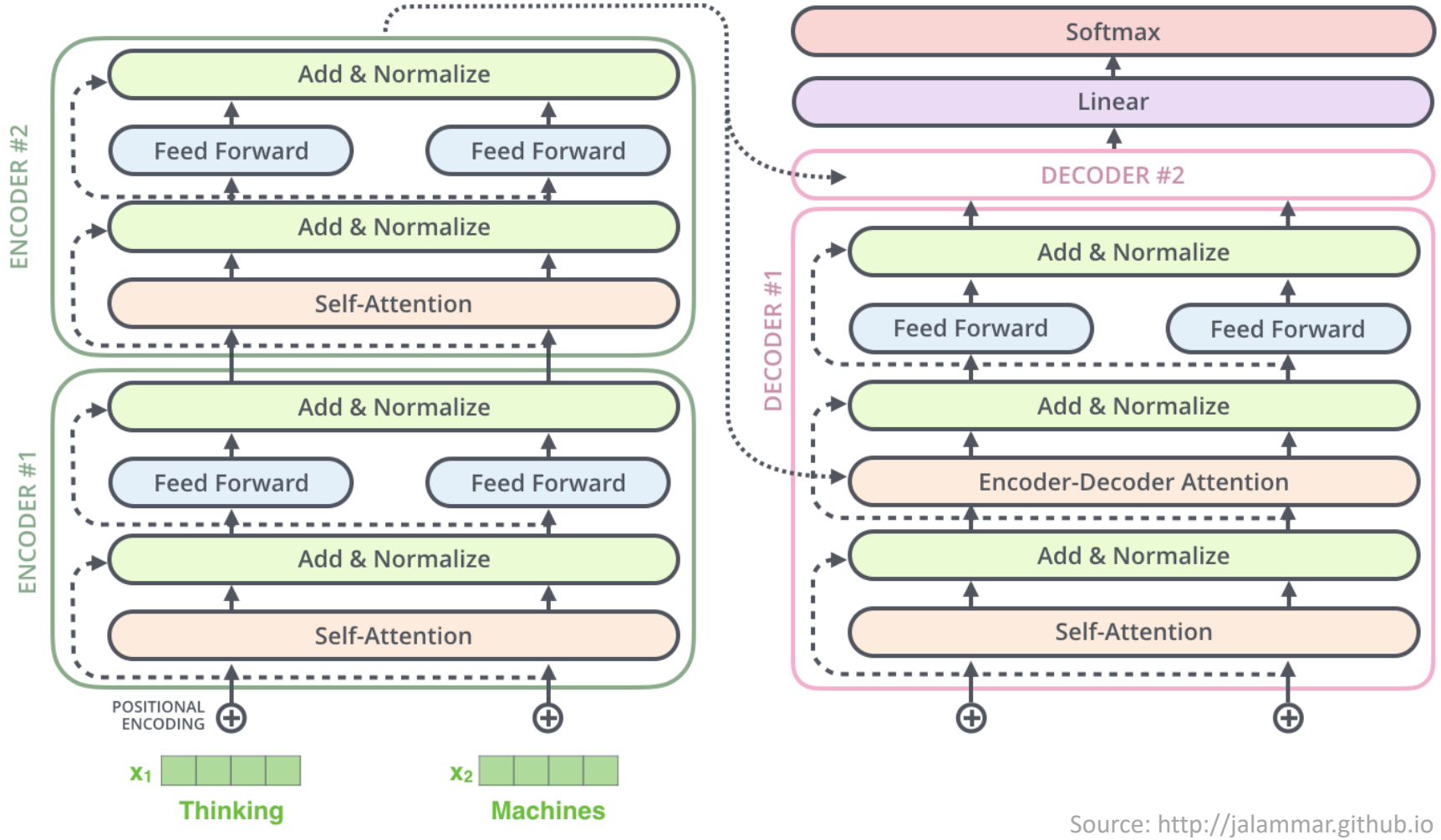
Source: http://jalammar.github.io

Recap: Transformer



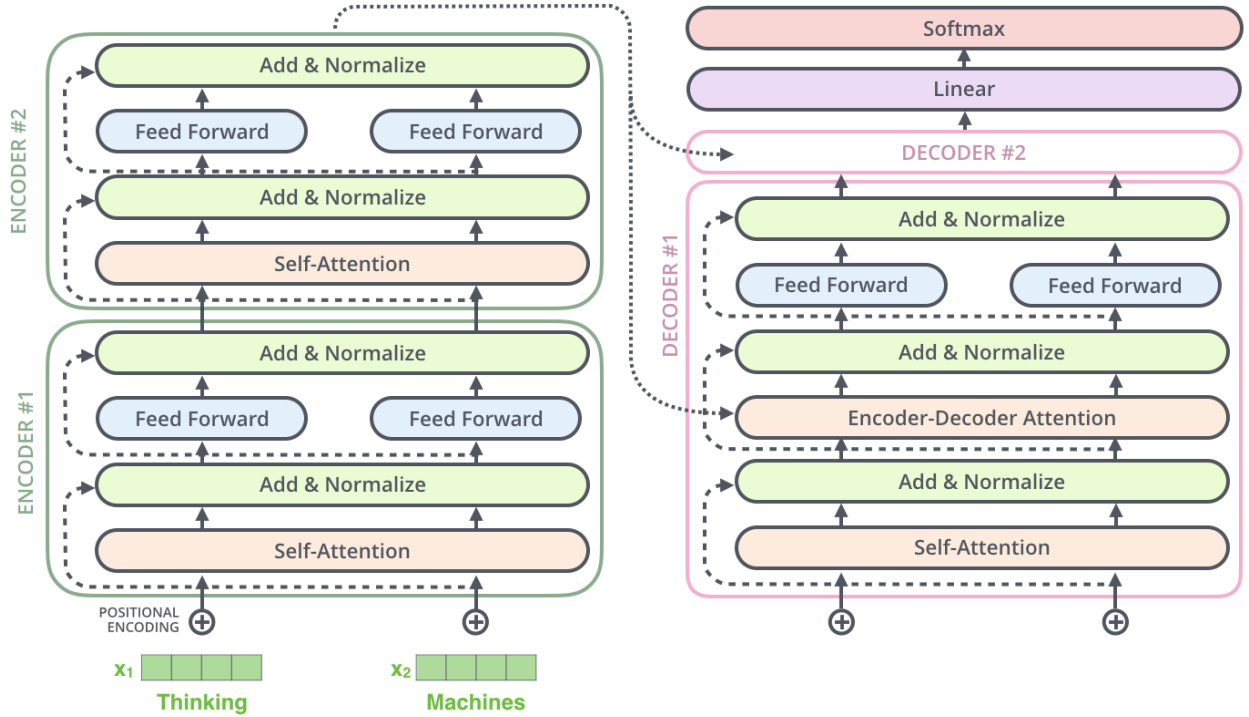
Source: http://jalammar.github.io

Recap: Transformer



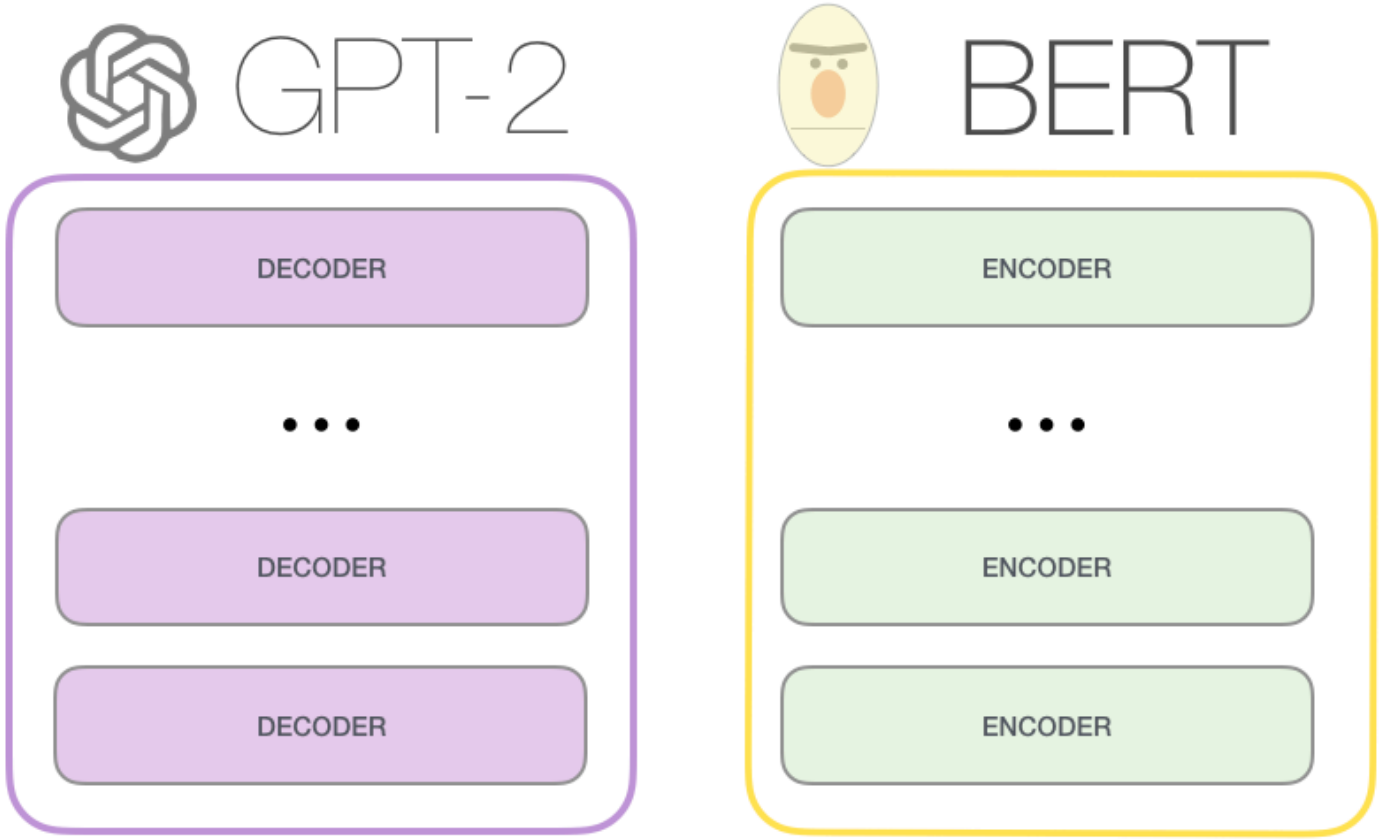
Source: http://jalammar.github.io

Recap: Transformer



$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Recap: GPT vs BERT



Source: http://jalammar.github.io

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Why GPT-2 rocks?

- Generates coherent paragraphs of texts
- SOTA performance on various language modeling benchmarks
- *not so bad* performance on several tasks with NO supervised training

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 <…>

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Approach

- Learning to perform single task: $p(output|input)$
- Learning to perform multiple tasks: $p(output|input, task)$

- But language provides ways to specify *output, input and task* in one sequence:

```
(translate to french, english text, french text).
(answer the question, document, question, answer).
```

- MQAN (McCann et al., 2018) – single model to perform different tasks with such a format

Approach

- But language provides ways to specify *output, input and task* in one sequence:

```
(translate to french, english text, french text).
(answer the question, document, question, answer).
```

- Supervised objective = unsupervised, but evaluated on a subset.
- Thus, global minimum of unsupervised obj. = GM of supervised.
- And we don’t need explicit supervision.

Training dataset

- Wikipedia, fiction books, news articles – single domain
- Common Crawl – data quality issues

- Solution: WebText
- WebText = outbound links from Reddit with $karma \geq 3$
- Over 8M documents
- No Wikipedia

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Deeper look inside GPT-2

- Input representation – Byte Pair Encoding
- BPE algorithm in a nutshell:
  - Split word to sequence of characters.
  - Joining the highest frequency pattern
  - Keeping doing previous step until it hit the pre-defined maximum number of sub-word of iterations.

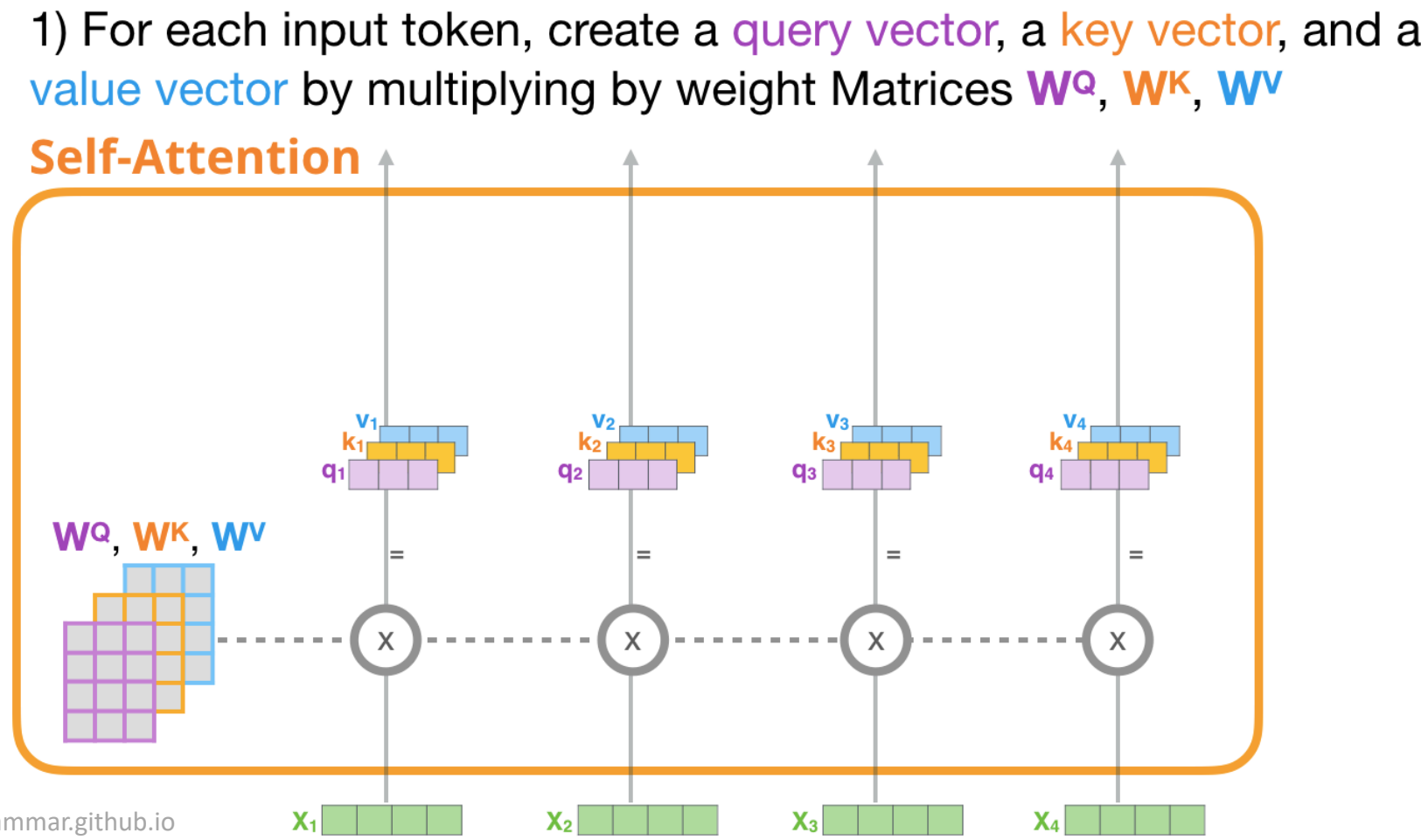
```
{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3}

{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w es t </w>': 6, 'w i d es t </w>': 3}

{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w est </w>': 6, 'w i d est </w>': 3}
```

Deeper look inside GPT-2

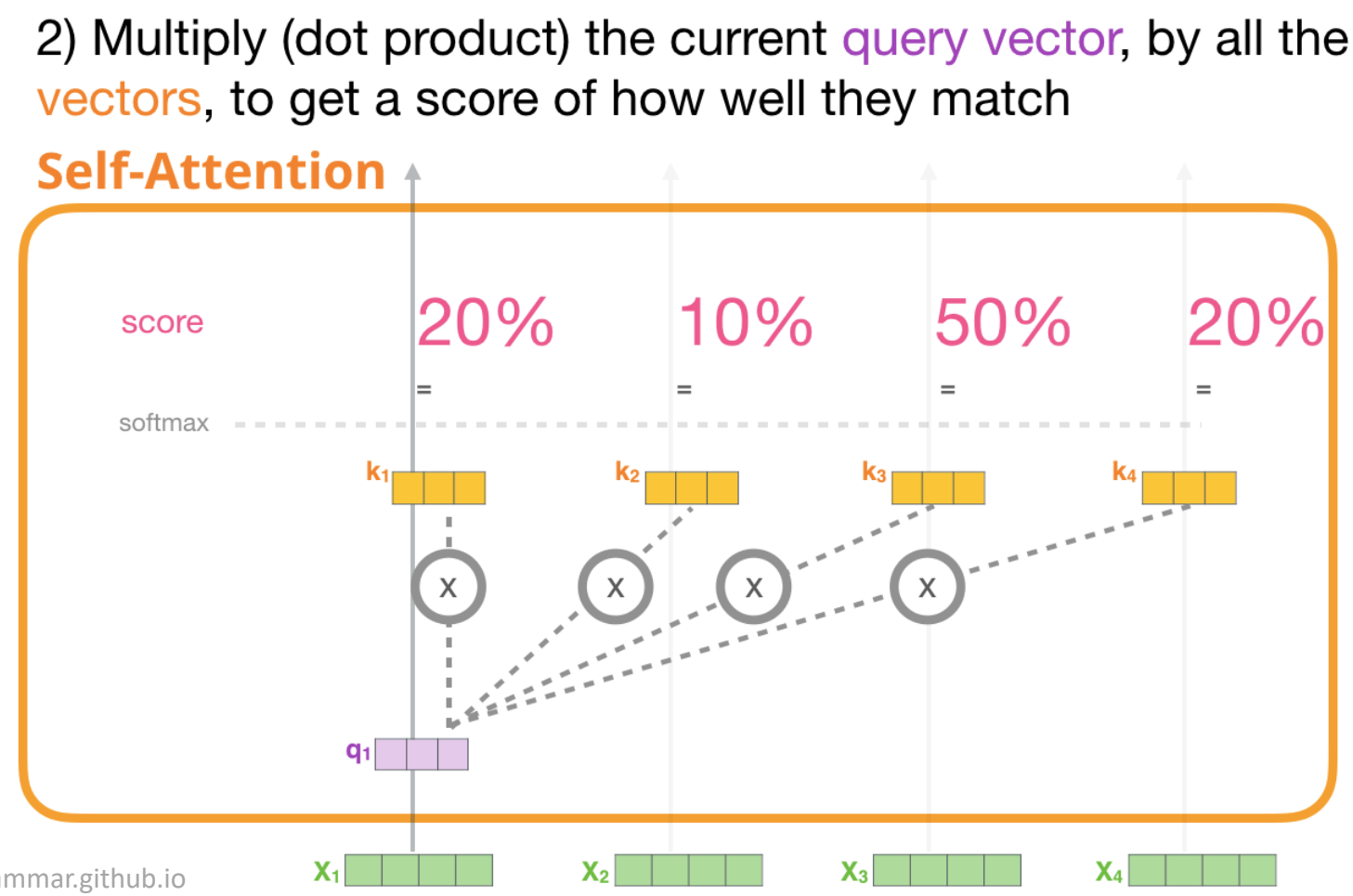
- Self-attention (no masking)



Source: http://jalammar.github.io

Deeper look inside GPT-2

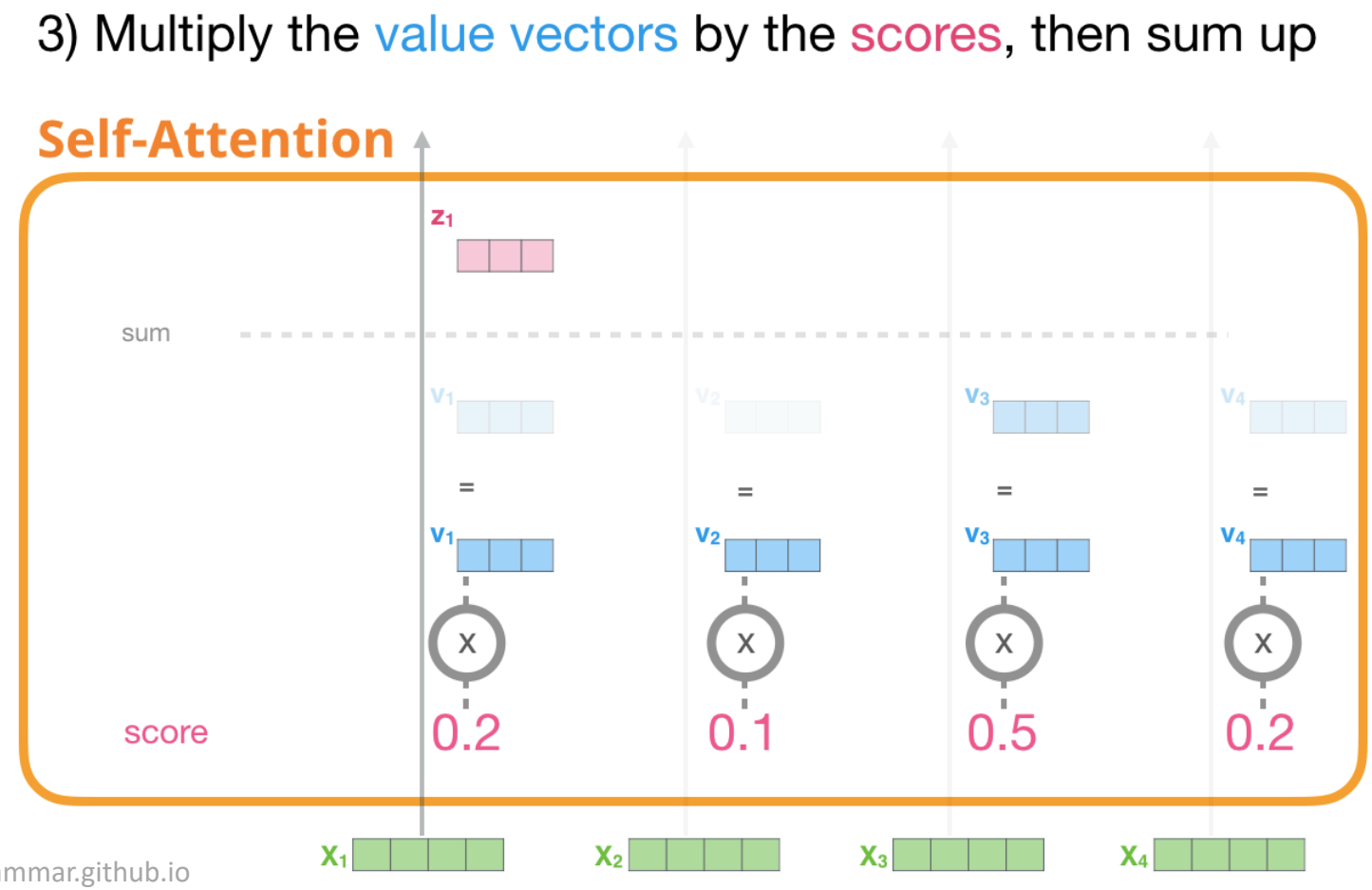
- Self-attention (no masking)



Source: http://jalammar.github.io

Deeper look inside GPT-2

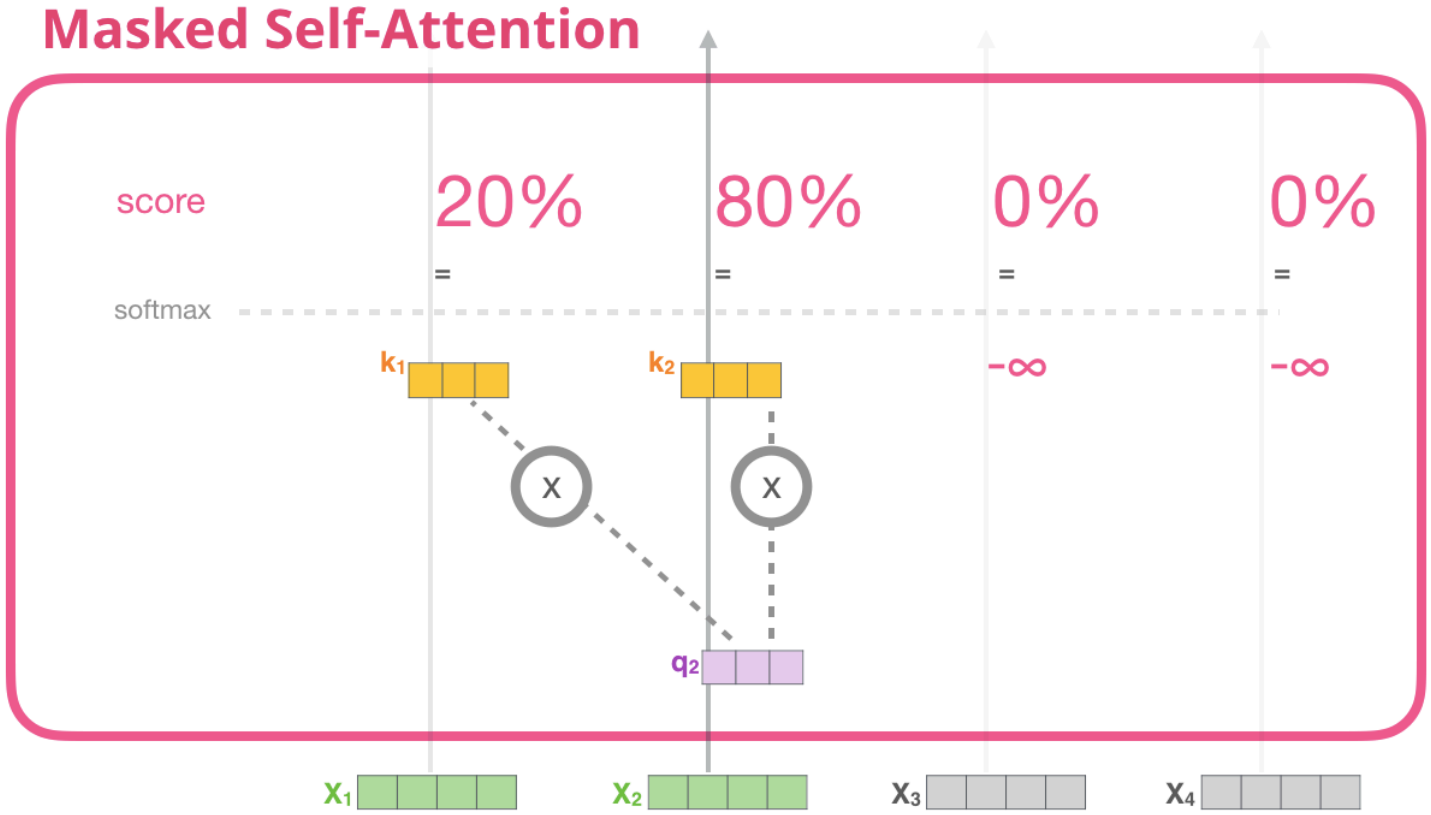
- Self-attention (no masking)



Source: http://jalammar.github.io

Deeper look inside GPT-2

- Self-attention (with masking)



Source: http://jalammar.github.io

Deeper look inside GPT-2

- Self-attention (with masking)

Masked Scores
(before softmax)

0.11	-inf	-inf	-inf
0.19	0.50	-inf	-inf
0.53	0.98	0.95	-inf
0.81	0.86	0.38	0.90

Softmax
(along rows)

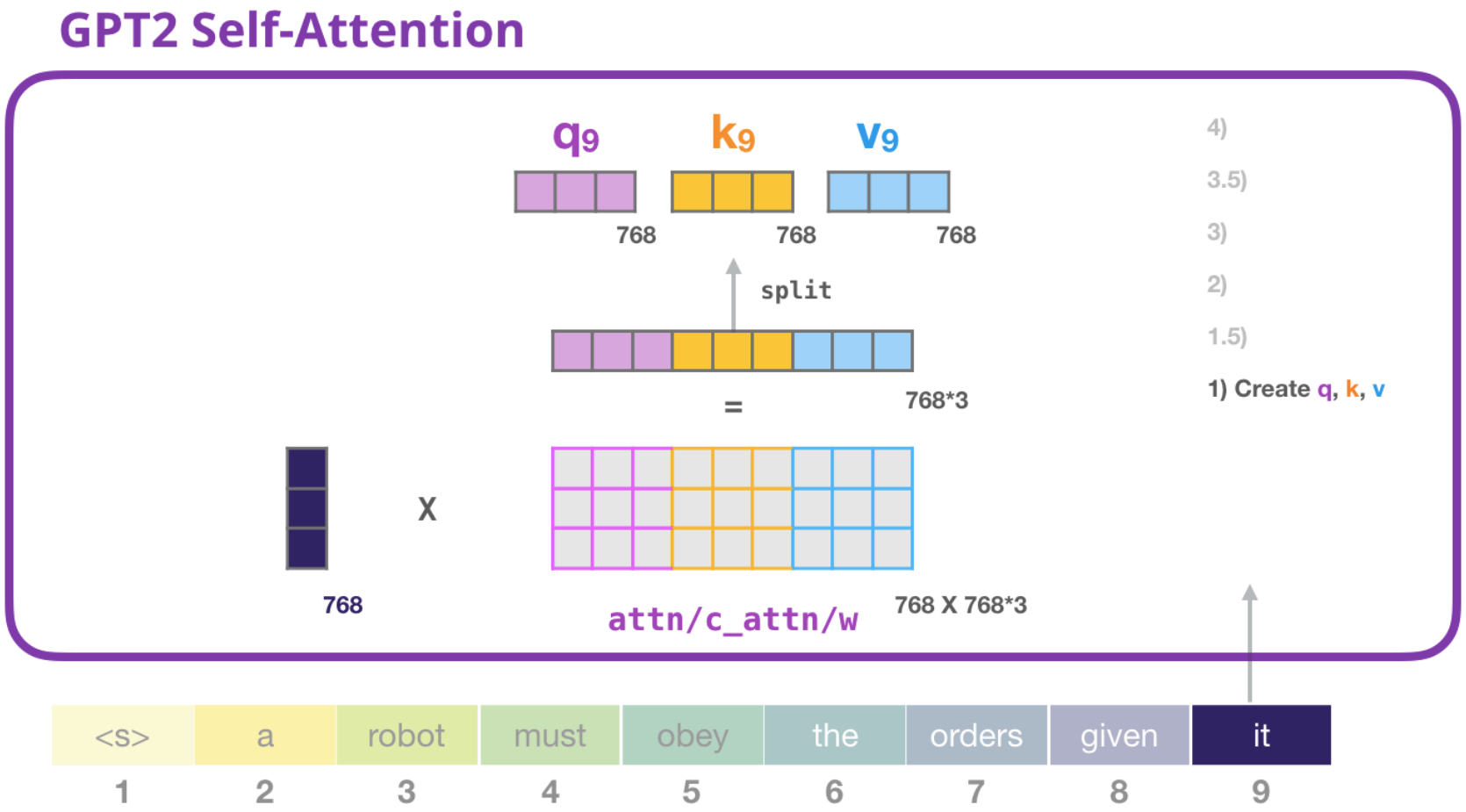
Scores

1	0	0	0
0.48	0.52	0	0
0.31	0.35	0.34	0
0.25	0.26	0.23	0.26

Source: http://jalammar.github.io

Deeper look inside GPT-2

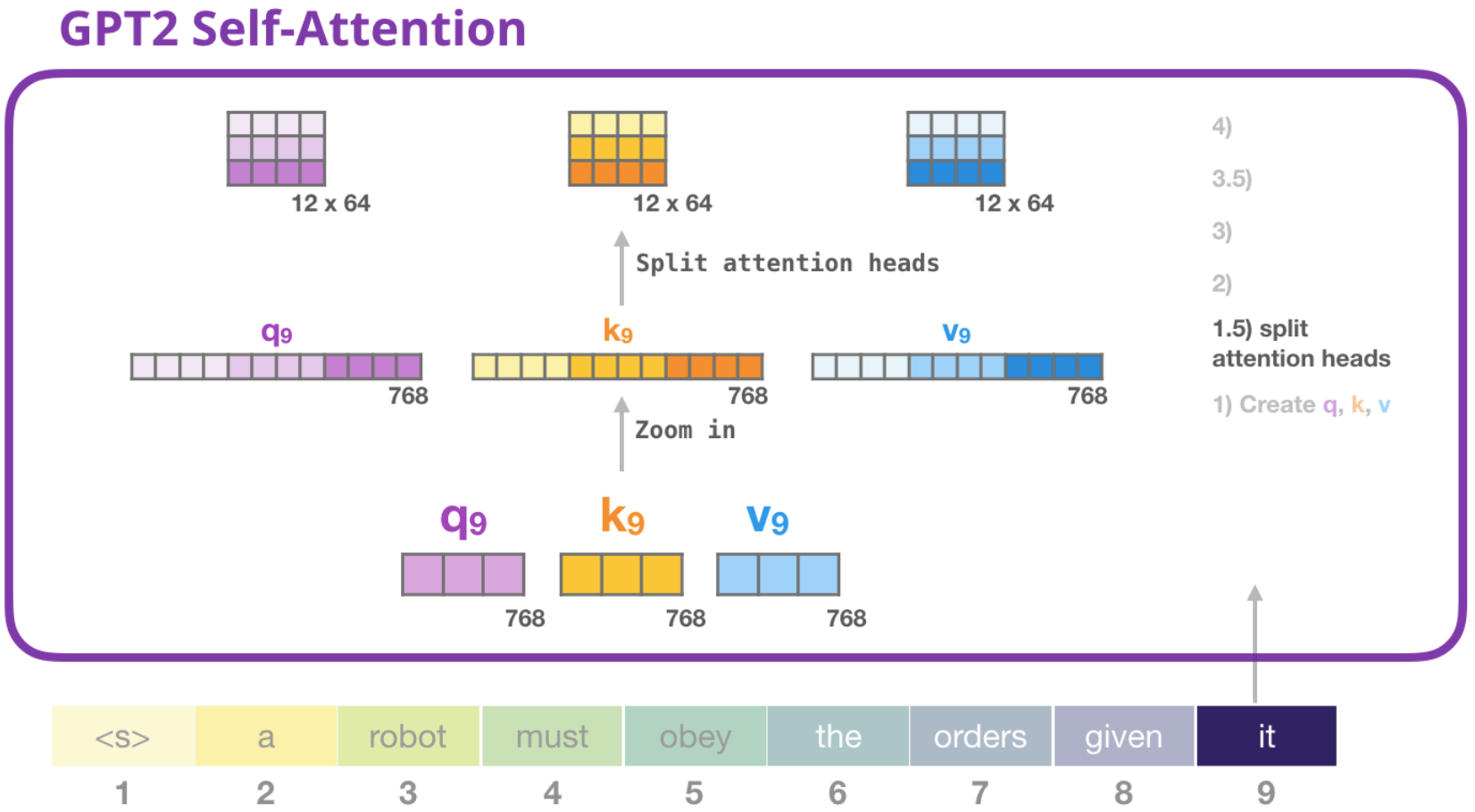
- Multi-head version



Source: http://jalammar.github.io

Deeper look inside GPT-2

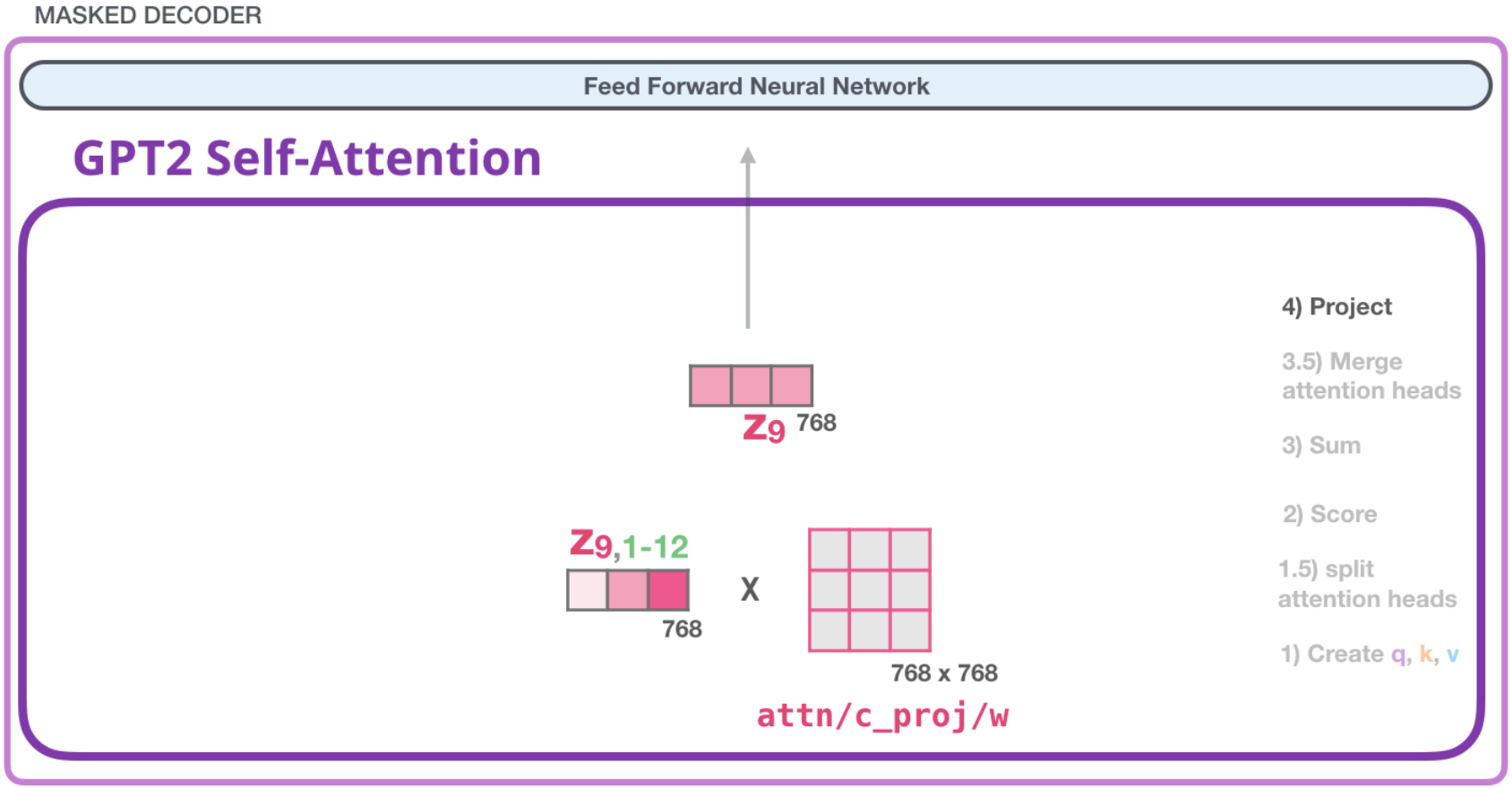
- Multi-head version



Source: http://jalammar.github.io

Deeper look inside GPT-2

- Multi-head version

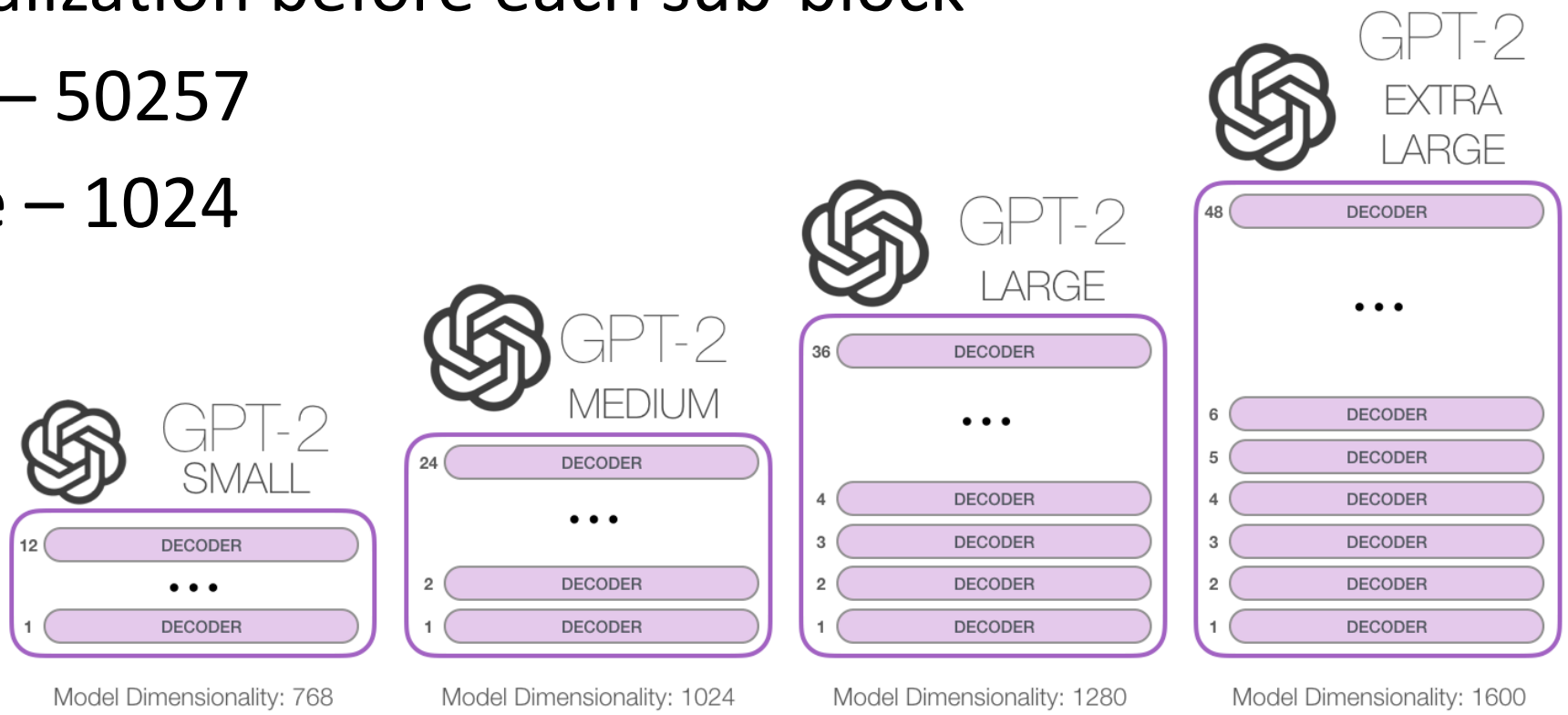


Source: http://jalammar.github.io

Deeper look inside GPT-2

Some extra facts:

- Positional encodings *(see "Attention Is All You Need" Vaswani et al.)*
- Layer normalization before each sub-block
- Vocabulary – 50257
- Context size – 1024



Source: http://jalammar.github.io

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Zero-shot task results

- Perplexity – a measure of how well the model predicts test data

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i | w_{i-k}, \ldots, w_{i-1})}$$

- Minimizing perplexity is the same as maximizing probability
- Better models have lower perplexity: they are less surprised by the test sample

Zero-shot task results

Children's Book Test

` The ogre is coming after us .
I saw him . '
` But where is he ?
I don't see him . '
` Over there .
He only looks about as tall as a needle . ‘
<…>
Then they both began to run as fast as they could I will get through it somehow , if I burrow underground , ' cried he , and very soon he and the XXXXX were on the other side .

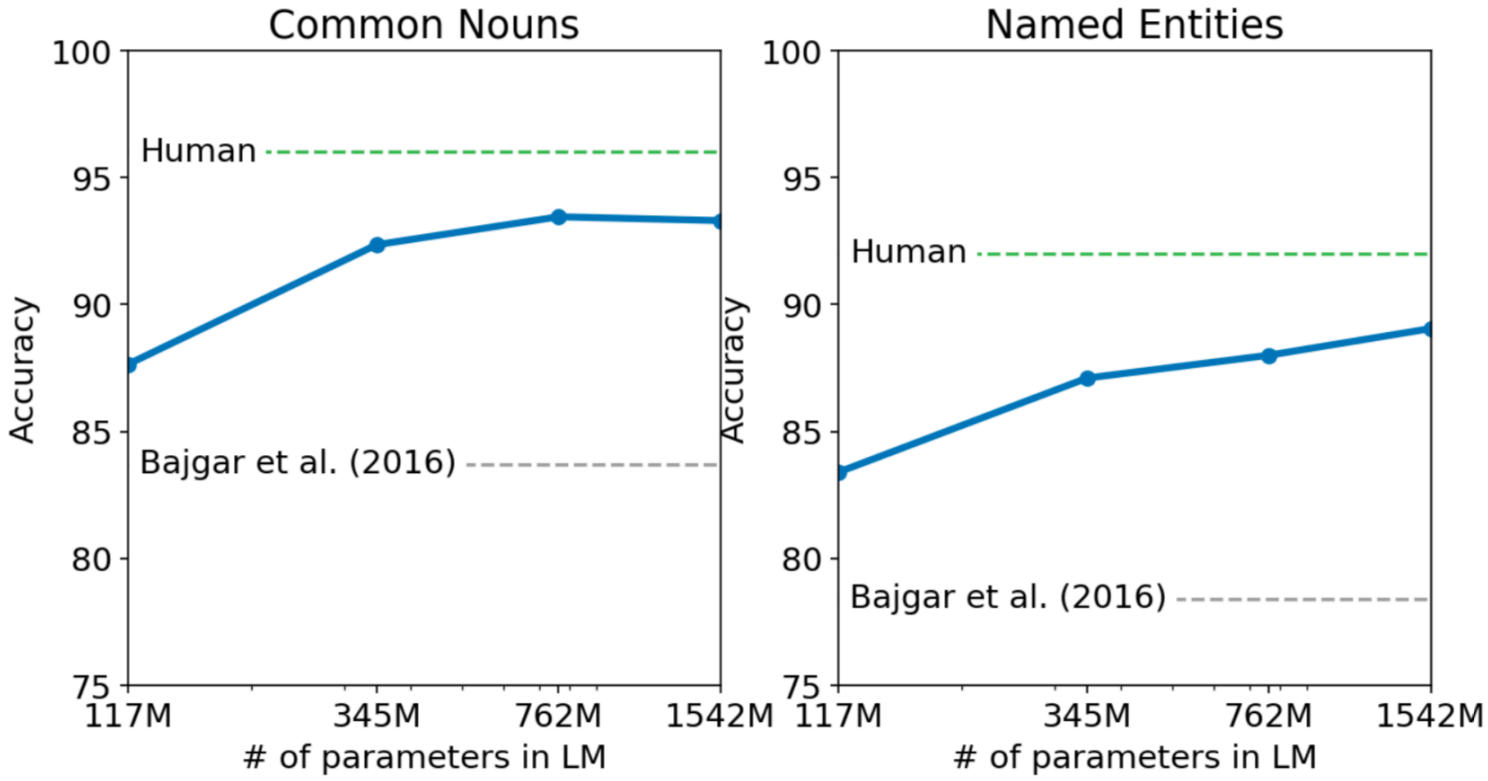
[correct answer]: dog
[answer canditates]: Cousin | cloak |
dog | maiden | mountain | needle | path | pin | side | steps

Four types of removed words:

- Common nouns
- Named entities
- Verbs
- Prepositions

Zero-shot task results

Children's Book Test



Source: Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

Zero-shot task results

LAMBADA

- Constraint on word to be final

	LAMBADA (PPL)	LAMBADA (ACC)
SOTA	99.8	59.23
117M	**35.13**	45.99
345M	**15.60**	55.48
762M	**10.87**	**60.12**
1542M	**8.63**	**63.24**

Context: They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.

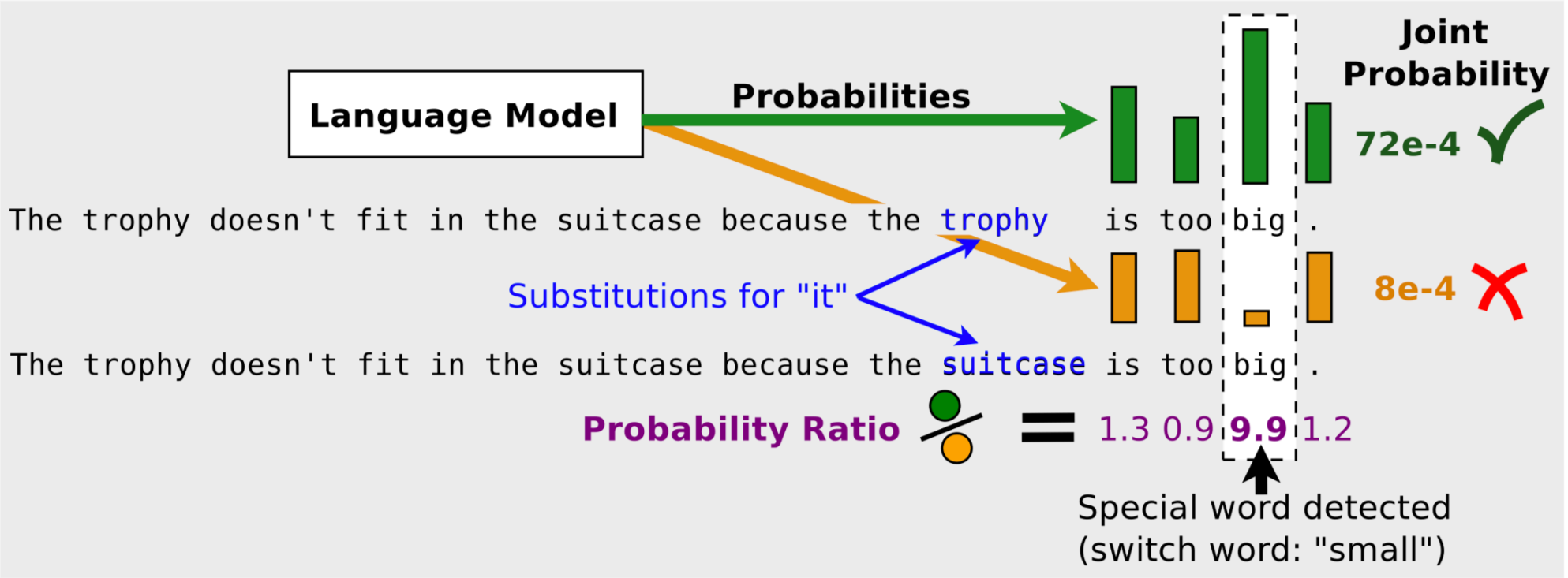
Target sentence: Aside from writing, I ’ve always loved _____.
Target word: dancing

Source: Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

Zero-shot task results

Winograd Schema Challenge

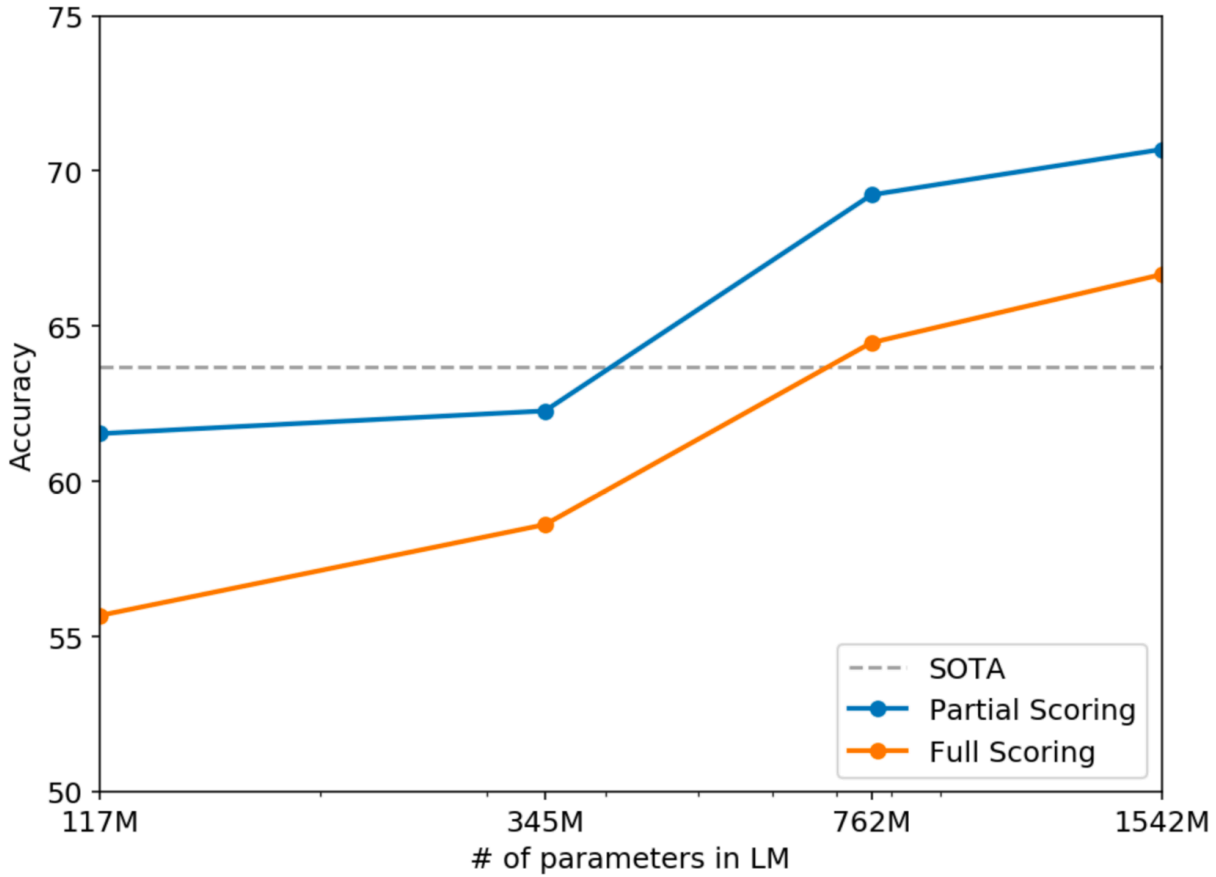
The trophy doesn't fit in the suitcase because it is too big. What is too big?
Answer 0: the trophy. Answer 1: the suitcase



Source: A Simple Method for Commonsense Reasoning (Trinh, Le, 2018)

Zero-shot task results

Winograd Schema Challenge



Source: Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

Zero-shot task results

The Conversation Question Answering dataset

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream <…>
Q: What was the theme?
A: "one world, one dream".
<…>
Q: Did they visit any notable landmarks?
A: Panathinaiko Stadium
Q: And did they climb any mountains?
A:

Target answers: *unknown* or *yes*
Model answer: Everest

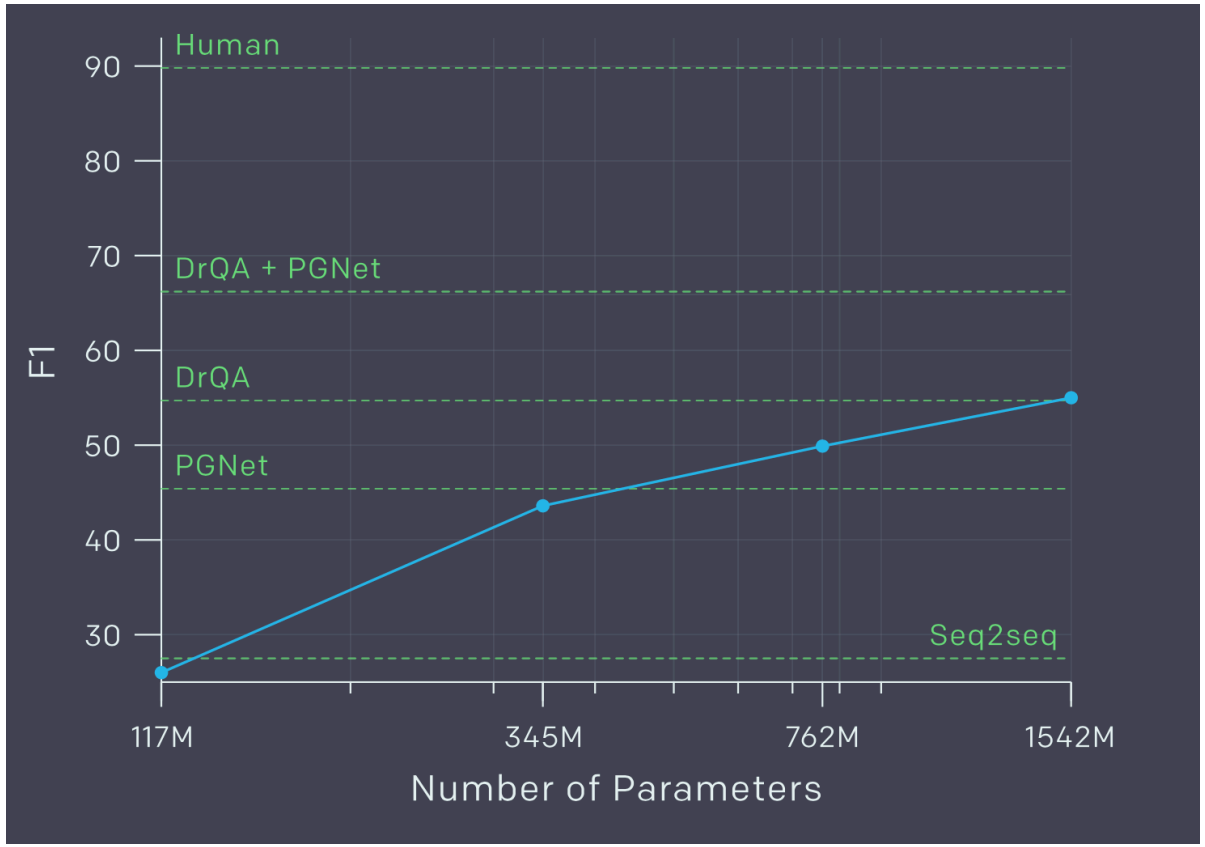
Source: https://openai.com/blog/better-language-models

Zero-shot task results

The Conversation Question Answering dataset

Error analysis:

- Simple retrieval based heuristics, e.g. *answer with a name from the document in response to a who question*



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

Zero-shot task results

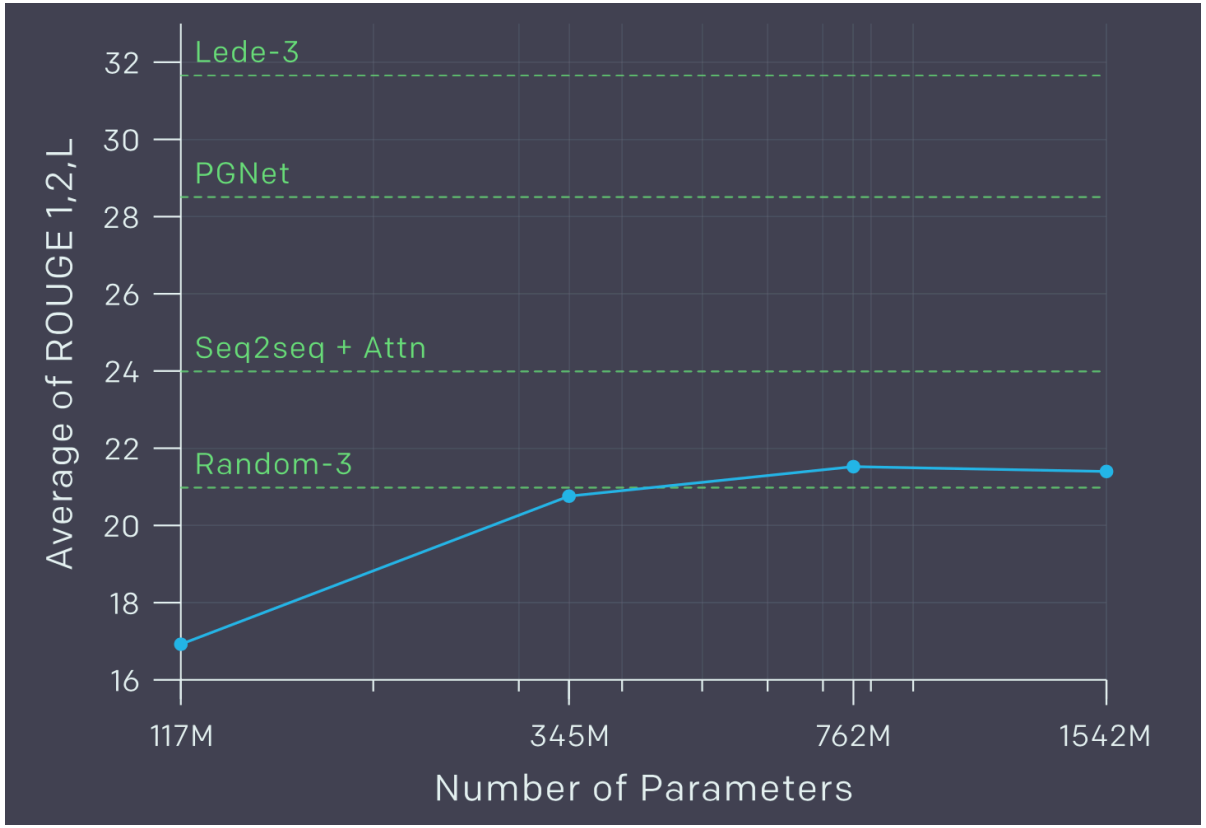
Summarization (CNN and Daily Mail dataset)

Approach:

- Text + “TL;DR:” (-6.4 R-AVG when no hint)
- Top-k random sampling, k=2
- 100 tokens are generated
- 3 sentences from these tokens

Cons:

- Focus on recent content
- Confuse specific details (e.g. how many cars were involved in a crash)



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

Zero-shot task results

Translation

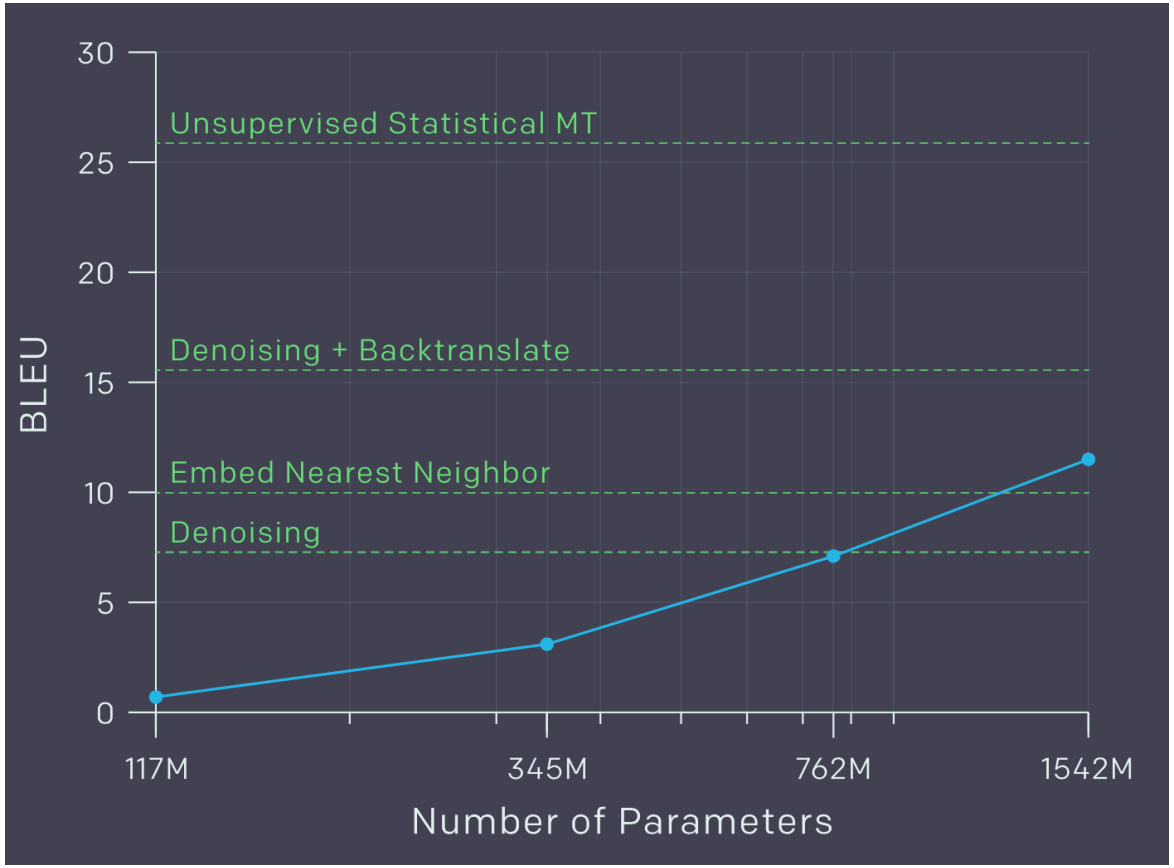
Approach:

- Pairs of <english sentence = french sentence> + “english sentence =”
- Greedy decoding

Facts:

- English-French: 5 BLEU
- French-English: 11.5 BLEU
- Non-English webpages was removed from WebText intentionally
- Only 10MB of French data left in dataset (≈500x smaller than the usual monolingual corpus)

WMT-14 Fr-En



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

Zero-shot task results

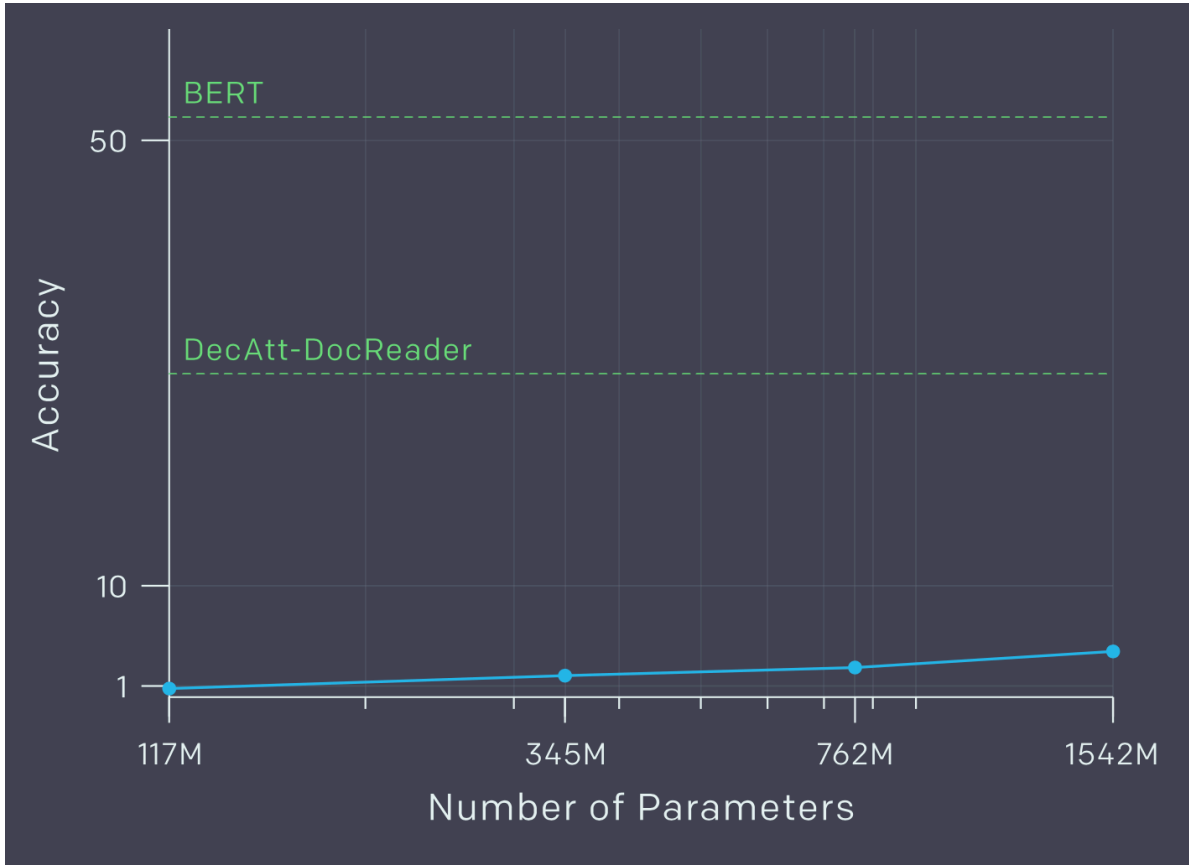
Question answering (Natural Questions dataset)

Approach:

- Pairs of <question, answer> + “question”

Facts:

- 63.1% accuracy on the 1% of questions the model is most confident in
- Smallest model accuracy < 1% (trivial baseline)



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

Plan

- Recap: LM task, Transformer, GPT vs BERT
- Why GPT-2 rocks?
- Approach and training dataset
- Deeper look inside GPT-2
- Zero-shot task results
- Generalization vs Memorization

Generalization vs Memorization

- Previously: CIFAR-10 – 3.3% overlap between train and test

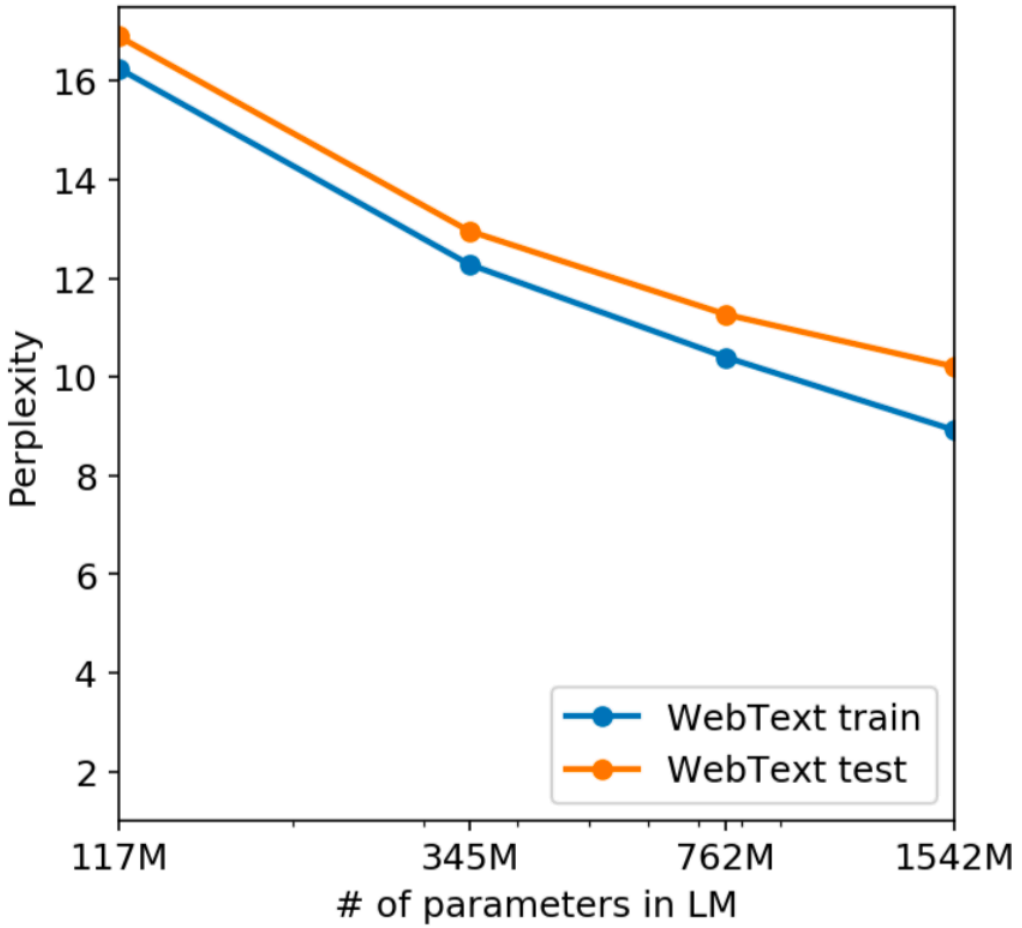
- Bloom filter for training 8-grams
- CoQA – 15% overlap, no actual questions or answers
- LAMBADA – 1.2% overlap
- Small but consistent benefit to results

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

Source: Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

Generalization vs Memorization



Source: Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

Summary

- Reformulated domain-specific tasks are a subset of general language modeling
- So the problem is to optimize the unsupervised objective to convergence
- When LM is trained on large and diverse dataset, it is able to perform well across many domains in a zero-shot setting

References

1. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D. & Sutskever, I. (2018), ’Language Models are Unsupervised Multitask Learners’.
2. Radford, A.; Narasimhan, K.; Salimans, T. & Sutskever, I. (2018), ‘Improving language understanding by generative pre-training’.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. & Polosukhin, I. (2017), ‘Attention is all you need’.
4. Better Language Models and Their Implications // URL: https://openai.com/blog/better-language-models
5. The Illustrated GPT-2 (Visualizing Transformer Language Models) // URL: http://jalammar.github.io/illustrated-gpt2