



Chapter 8

Word Representations^{part3}

Context

A circular watermark of the University of Würzburg seal, featuring a coat of arms with figures and text around the border.

Recall

Word Embeddings — Idea

Represent high dimensional **sparse vectors** (one-hot encoded words)
as low dimensional **dense vectors**

Then use these vectors as input to task specific model

- Vectors should capture the semantics of words
- Words that can be used interchangeably should have very similar vectors
- Word embeddings can be learned on large text corpora and used in other tasks

Biggest Problem of Word2Vec/...

He kicked the **bucket**.

I have yet to cross-off all the items on my **bucket** list.

The **bucket** was filled with water.

1. Words are sometimes ambiguous (e.g. bank, play, bucket, stick, ...)
2. Word meaning depends on the context it is used in
3. Word2Vec learns **one** vector per word

→ Make embeddings context-sensitive!

→ Embedding captures **current** context

→ Different embeddings for different contexts

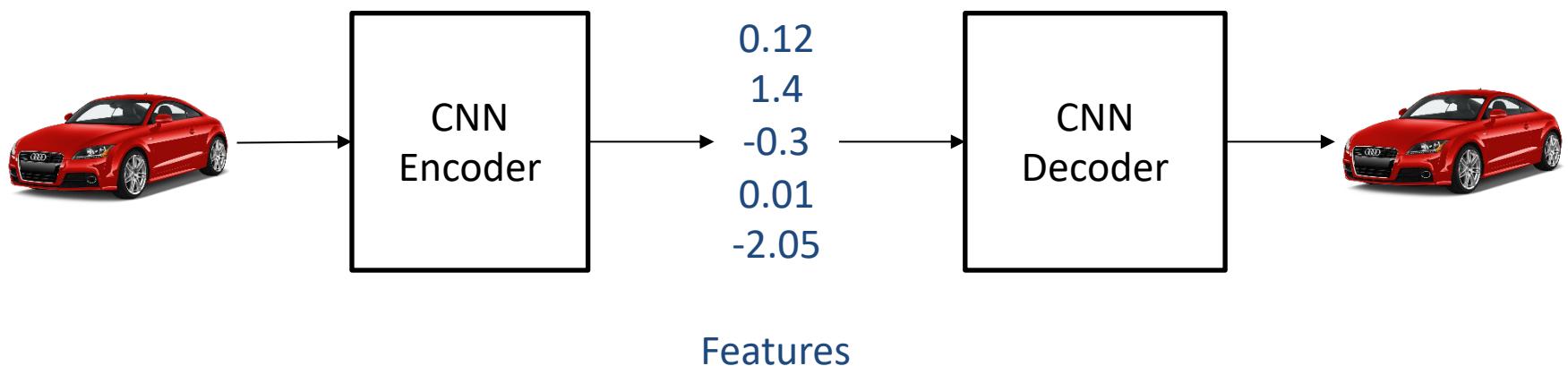


Why bother with Word Embeddings?

Reuse weights from pretrained model from another task!
→ Transfer Learning

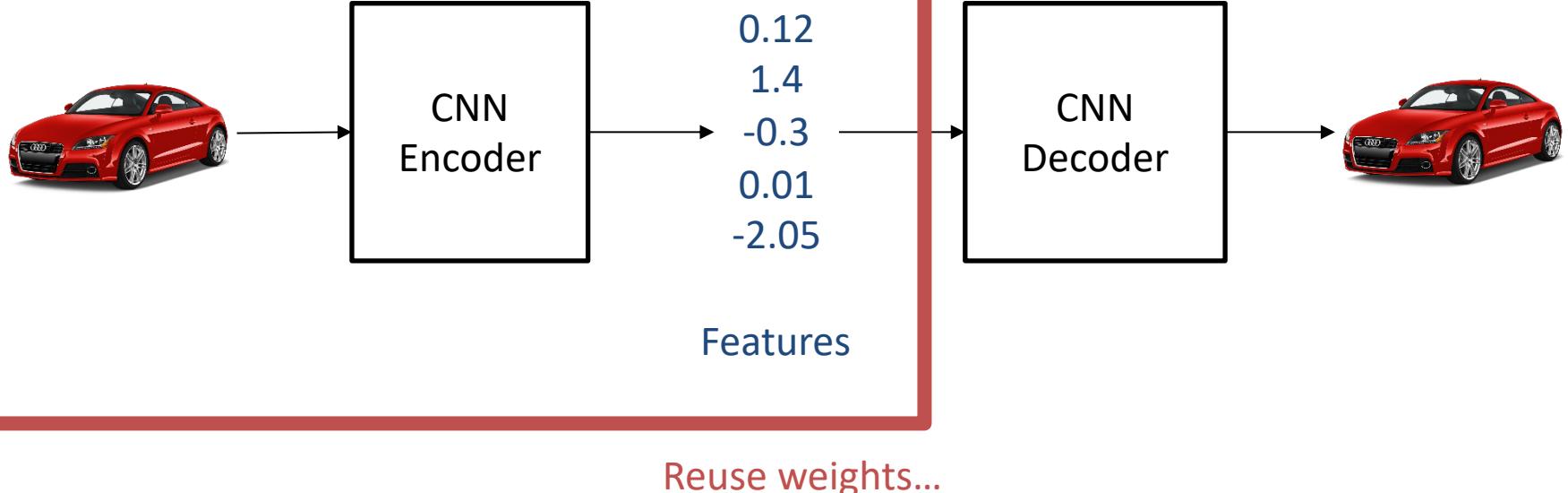
Transfer Learning (in Computer Vision)

Learn on one task/dataset,
then transfer to another task/dataset



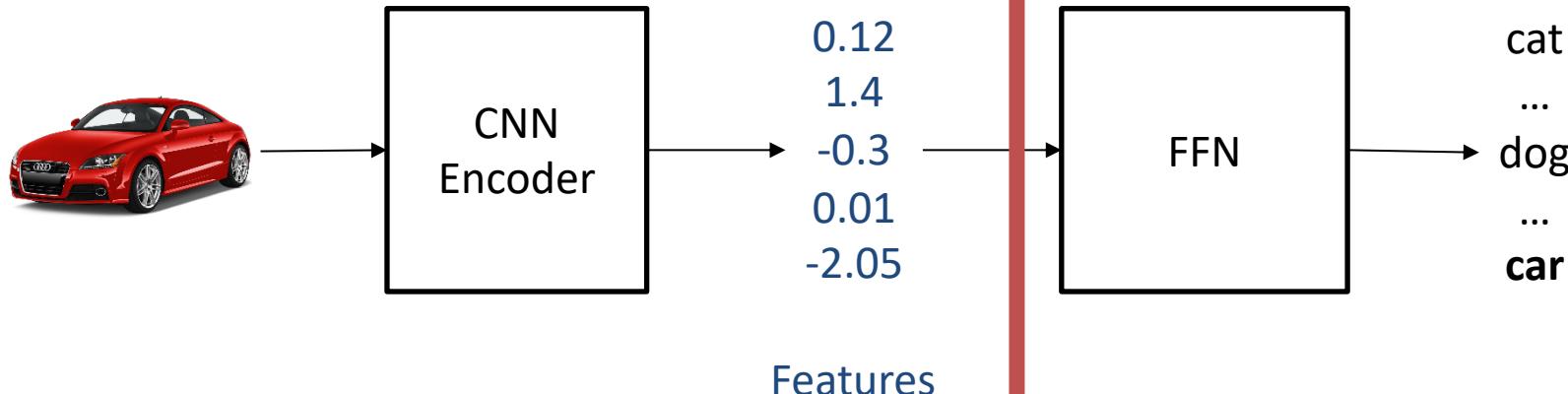
Transfer Learning (in Computer Vision)

Learn on one task/dataset,
then transfer to another task/dataset



Transfer Learning (in Computer Vision)

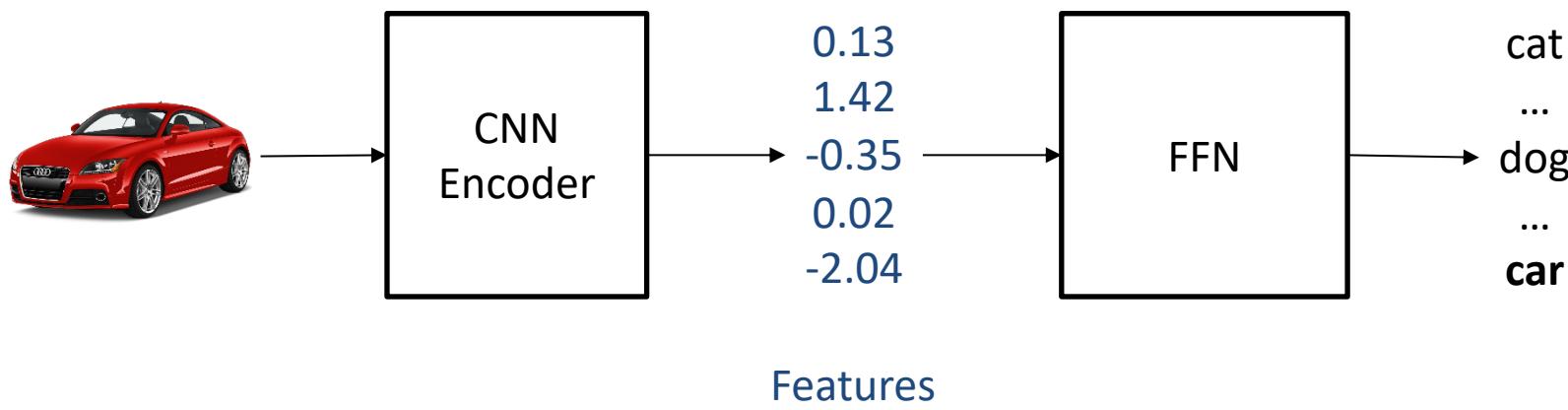
Learn on one task/dataset,
then transfer to another task/dataset



Reuse weights...in another task!

Transfer Learning (in Computer Vision)

Learn on one task/dataset,
then transfer to another task/dataset



Fine-tune complete model!



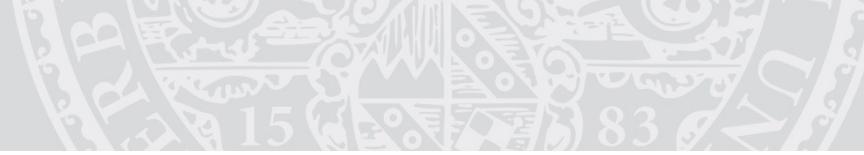
Transfer Learning in NLP

- Pretrain a model on a task with a big amount of data
- Modify model and reuse trained weights to solve a new task
- Better performance on tasks with less data
 - Model does not have to learn everything from scratch (syntax, word meaning, ...)
 - Pretraining task and target task should be similar → better performance
- Big theme in NLP for the last few years!



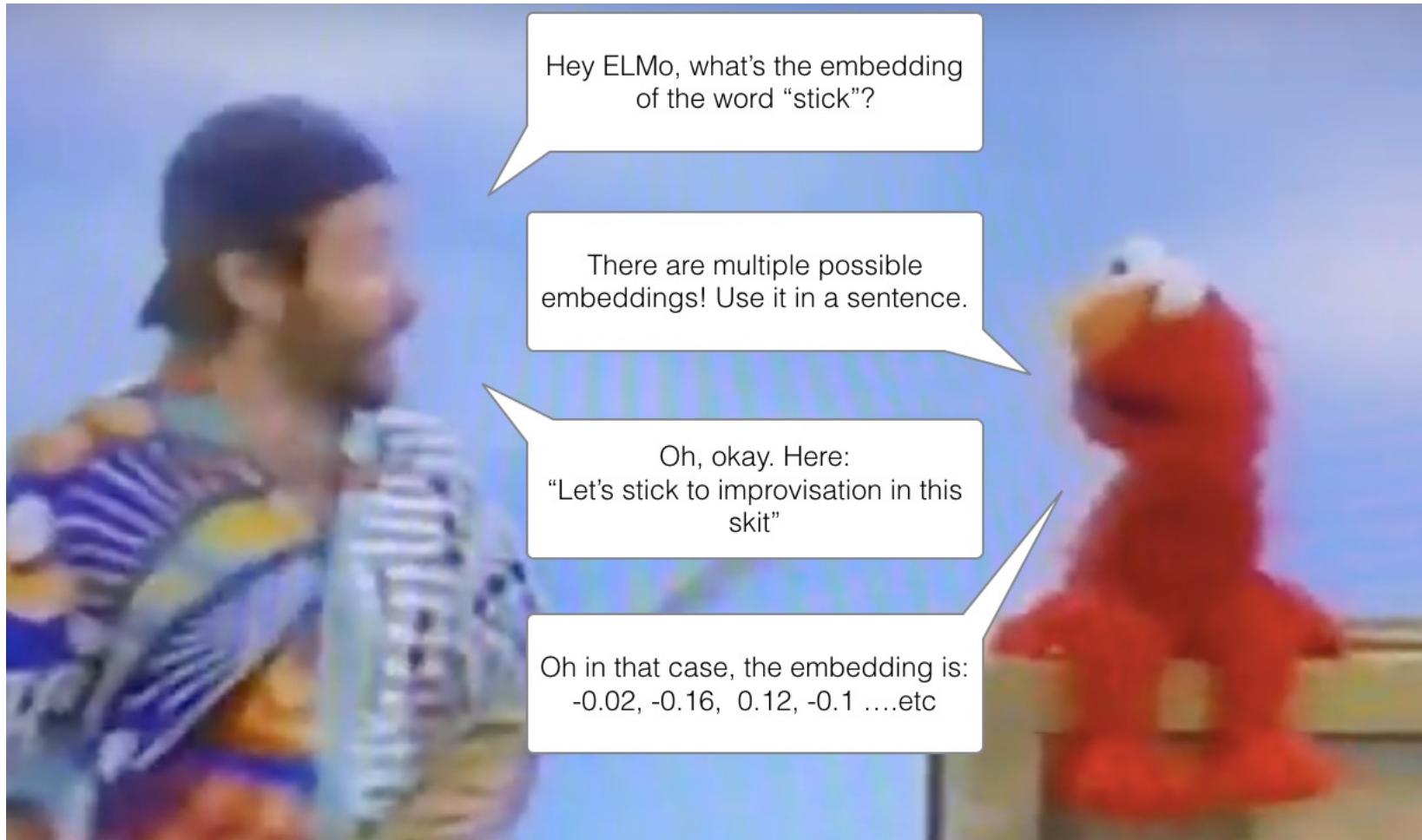
Transfer Learning in NLP

- What task to use?
- **Language Modelling** has basically unlimited training data!
- **Language Modelling** is also very general!
- **Language Modelling** has been shown to work well for embeddings!



One important part of Language Models is capturing context...

Enter ELMo!

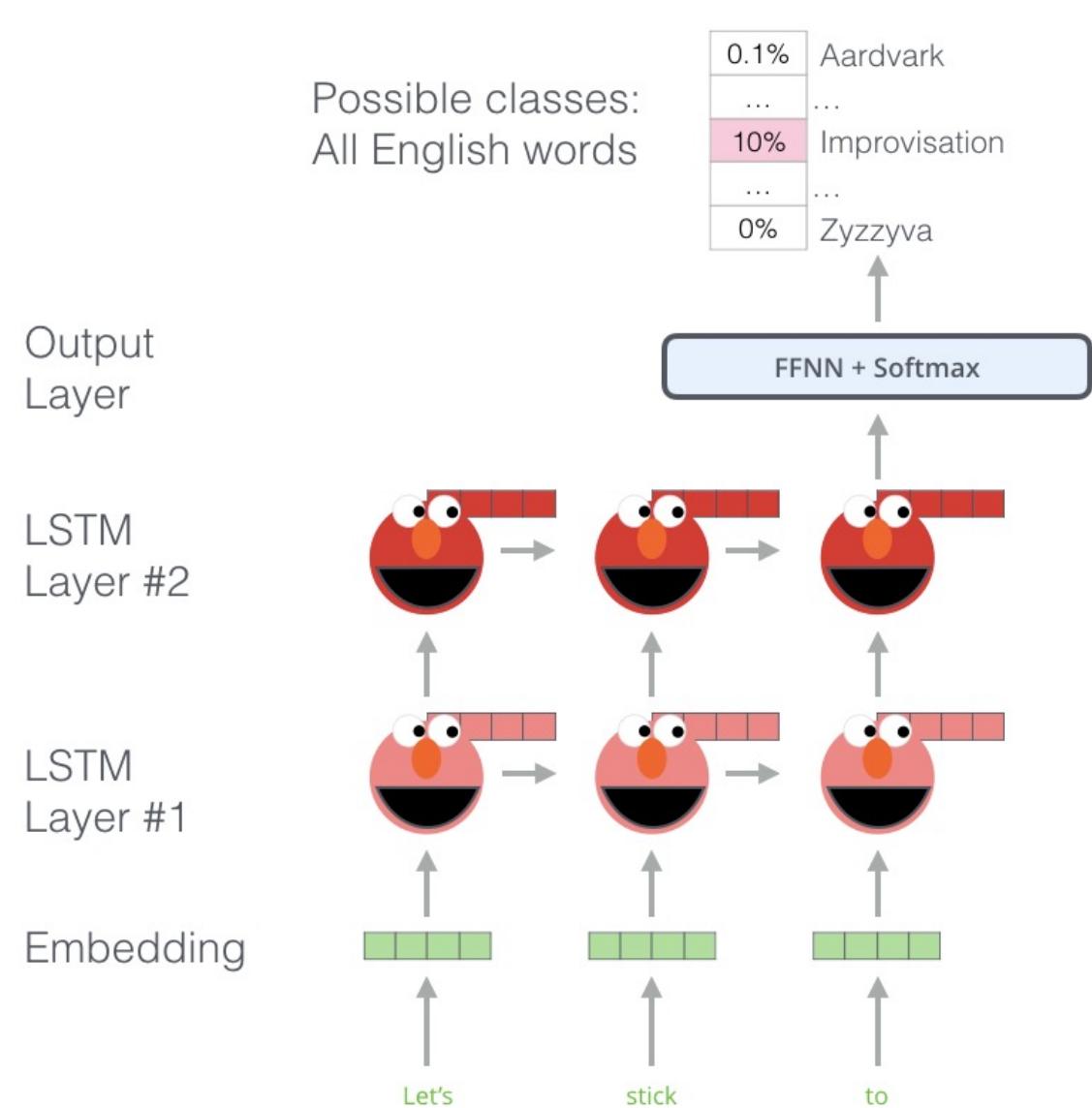


<http://jalammar.github.io/illustrated-bert/> and

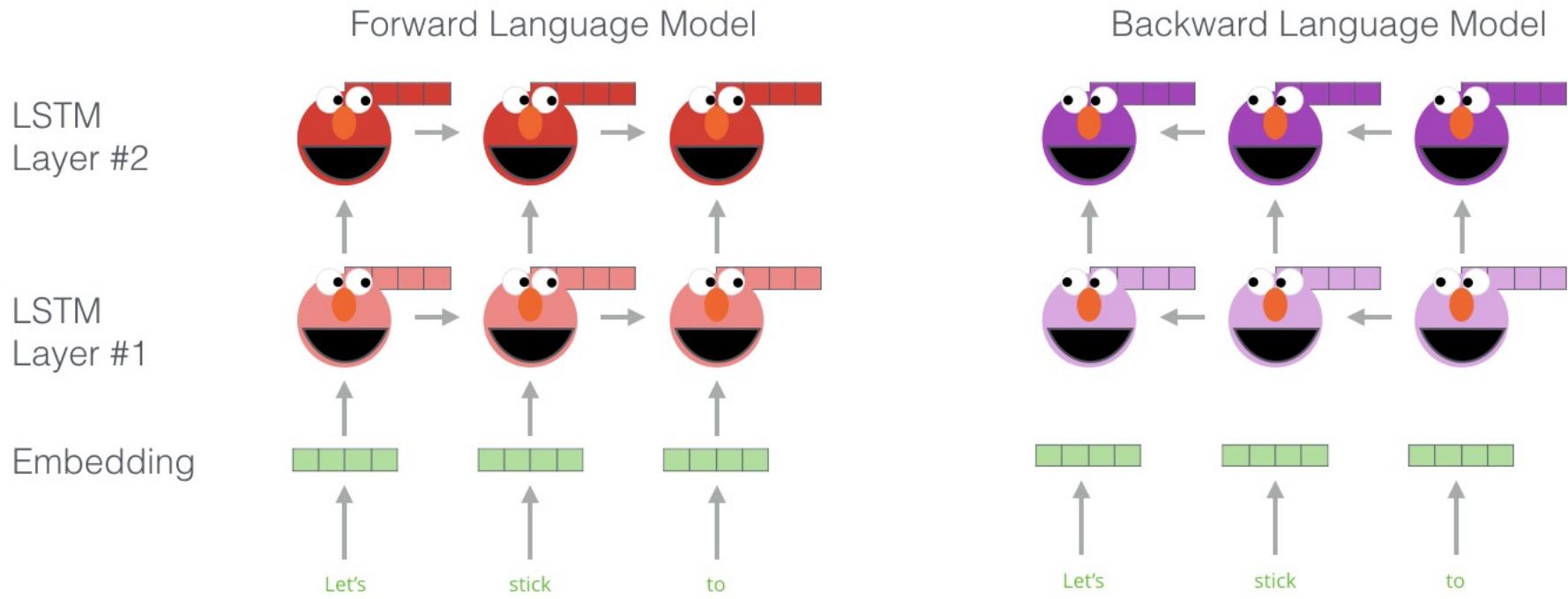
Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).

ELMo — Idea

- Train a **language model**: Predict next word given previous text
- Model learns syntax, grammar and word relations
- Meaning of each word is captured in hidden states



ELMo – Step 1



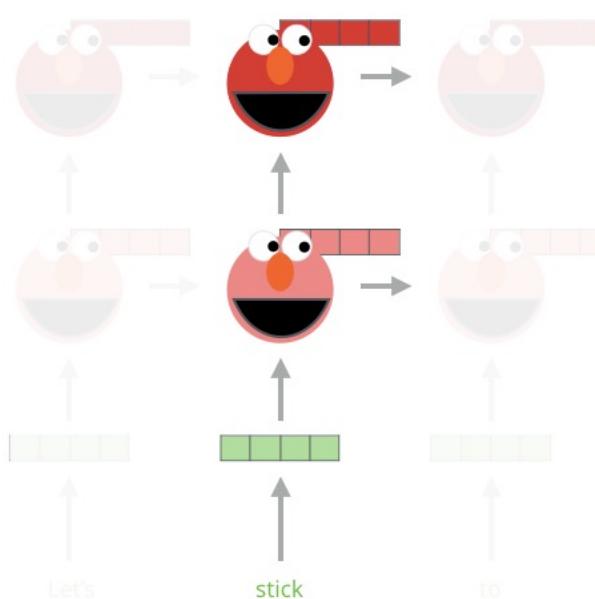
- Feed sentence through a **bidirectional language model**
 - Captures both directions of the sentence

ELMo – Step 2

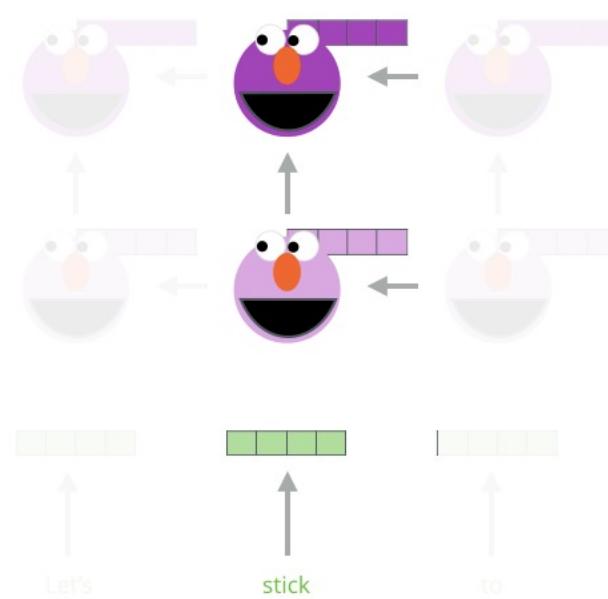
1- Concatenate hidden layers



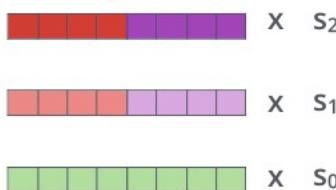
Forward Language Model



Backward Language Model



2- Multiply each vector by a weight based on the task

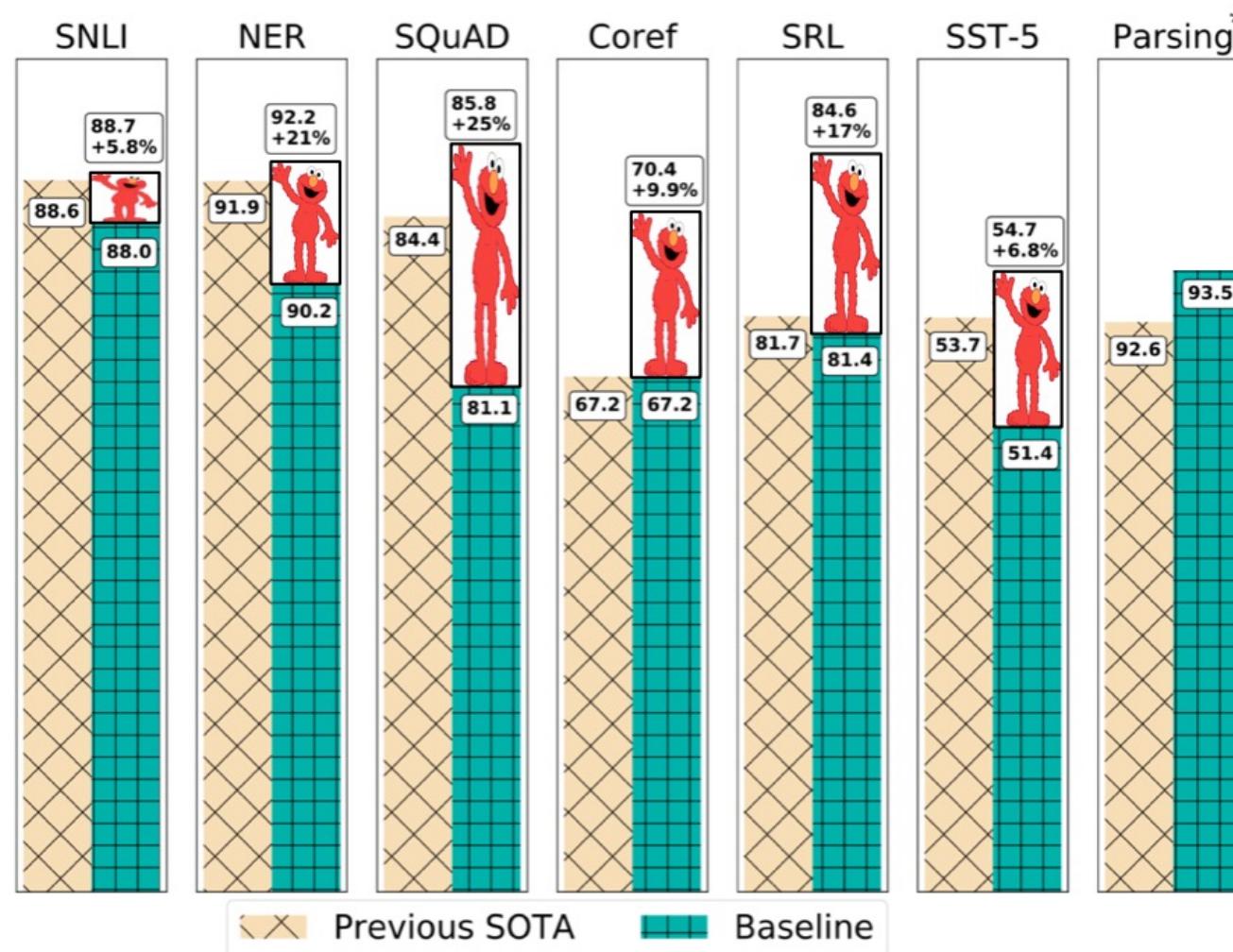


3- Sum the (now weighted) vectors



ELMo embedding of "stick" for this task in this context

ELMo — Results



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

ELMo — Pros and Cons

Pros

- Captures context
- Better performance

Cons

- No simple table lookup → load model and compute embedding

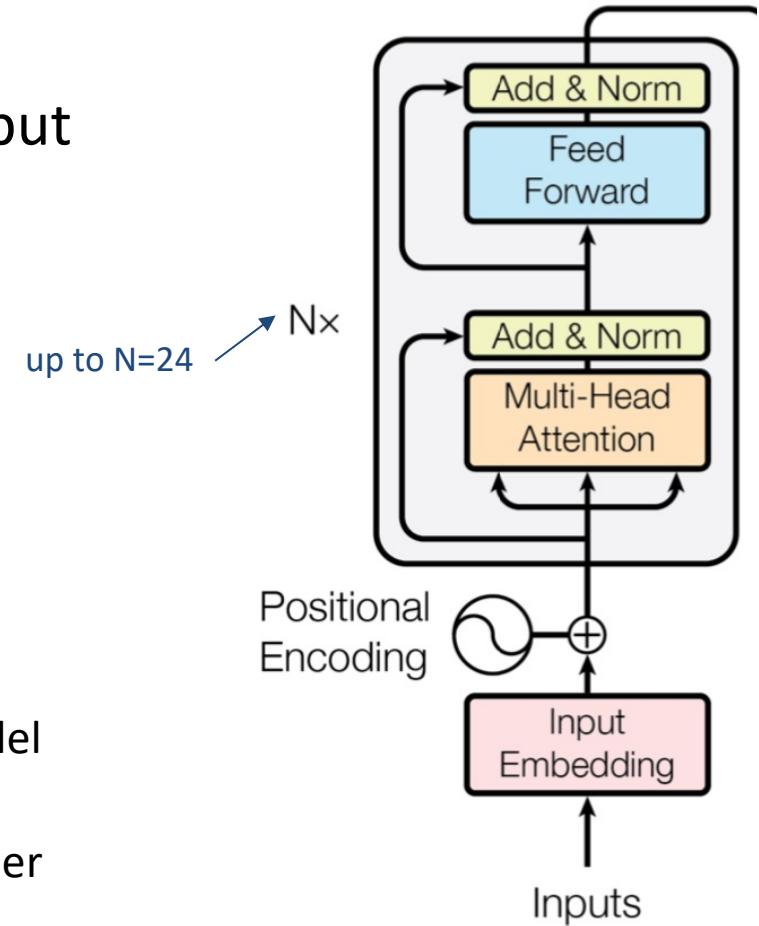


What could work better than LSTMs?



BERT – Idea

- Train a **Transformer** encoder
- Feed whole sentence to network but mask out words
 - Get bidirectionality with one model
- Train model on multiple tasks for which a big amount of data exists:
 - Predict masked word („masked language model“)
 - Randomly replace words and let the model predict the correct word
 - Ask model if a sentence follows on another sentence



BERT — Masked Language Model

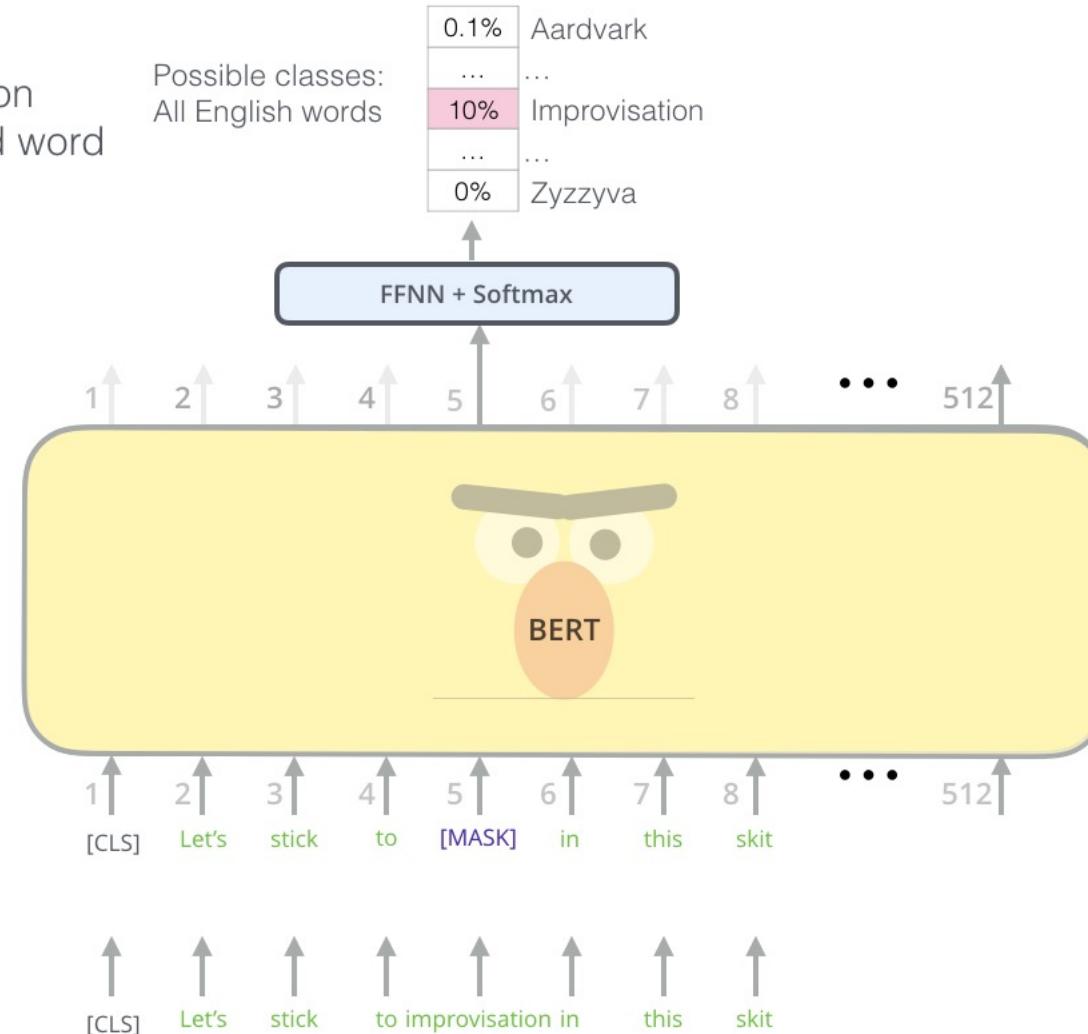
Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zzyzyva

FFNN + Softmax

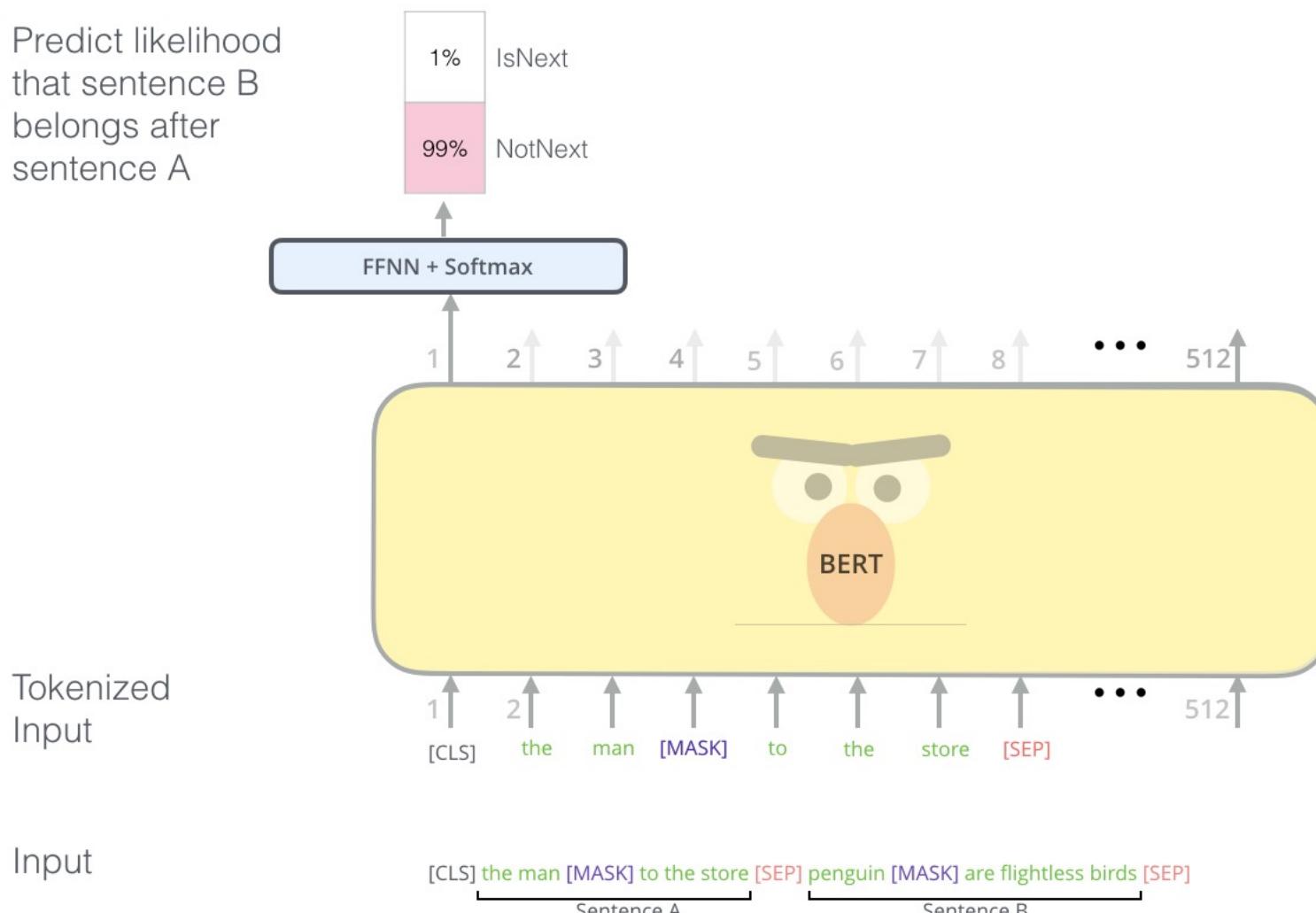
Randomly mask 15% of tokens



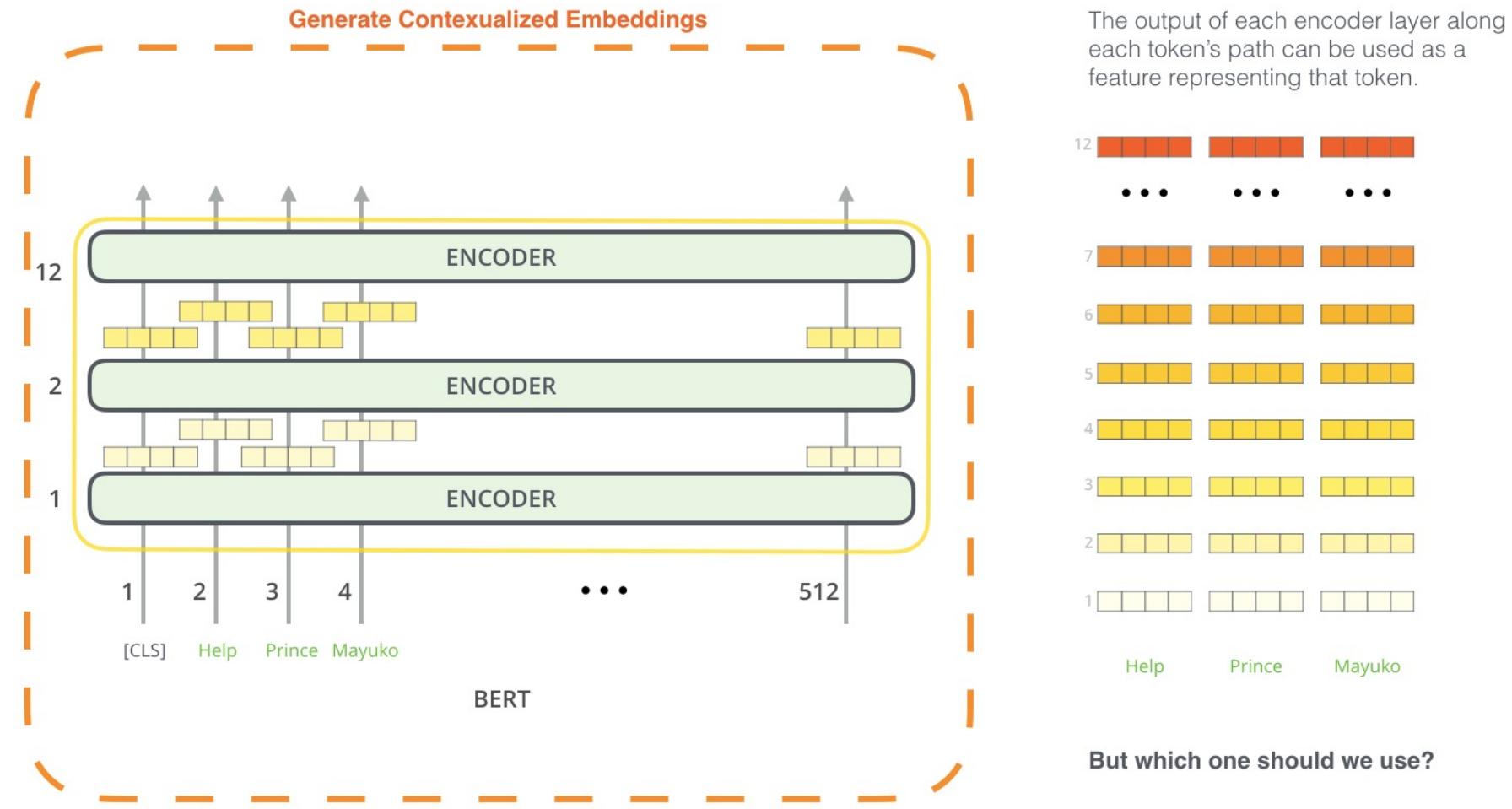
Input

BERT — Two Sentence Task

Predict likelihood
that sentence B
belongs after
sentence A



BERT — Contextualised Word Embeddings



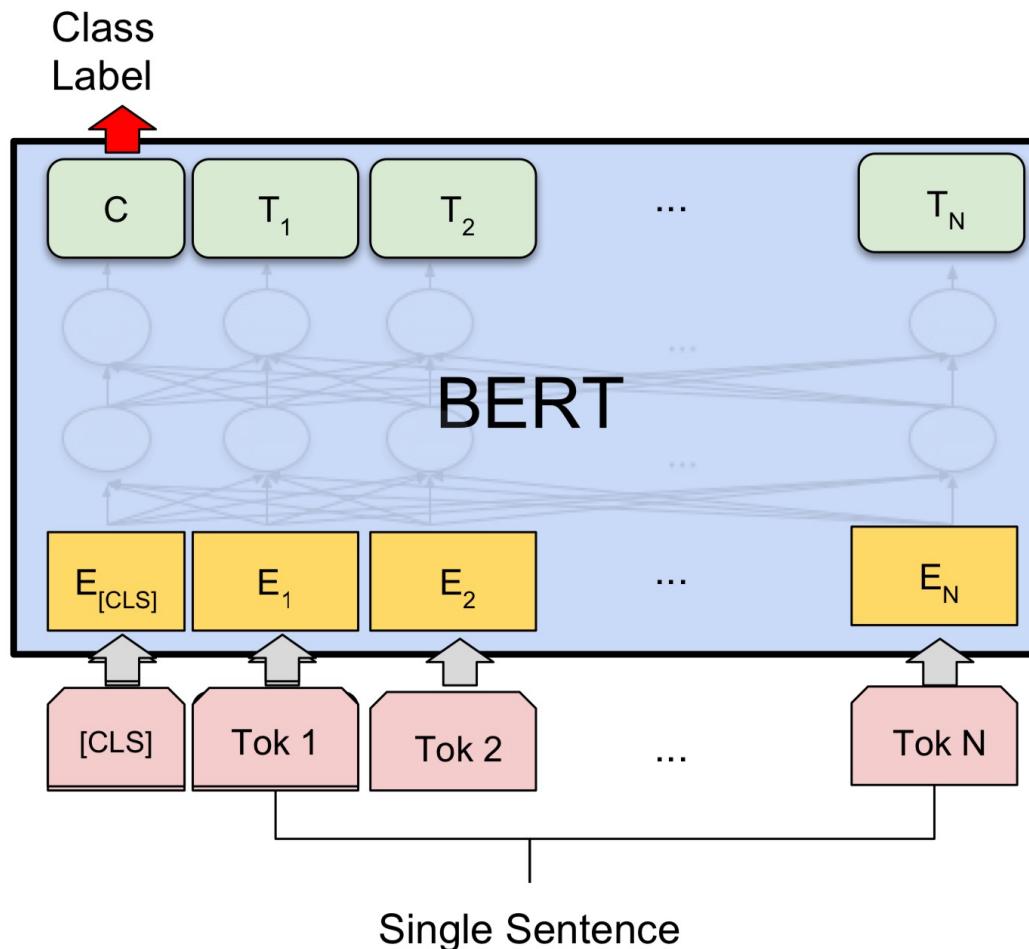
BERT — Contextualised Word Embeddings

What is the best contextualized embedding for “**Help**” in that context?

For named-entity recognition task CoNLL-2003 NER

		Dev F1 Score
12		91.0
• • •		
7		94.9
6		95.5
5		
4		
3		95.6
2		
1		95.9
Help		
First Layer	Embedding	
Last Hidden Layer		
Sum All 12 Layers		
Second-to-Last Hidden Layer		
Sum Last Four Hidden		
Concat Last Four Hidden		96.1

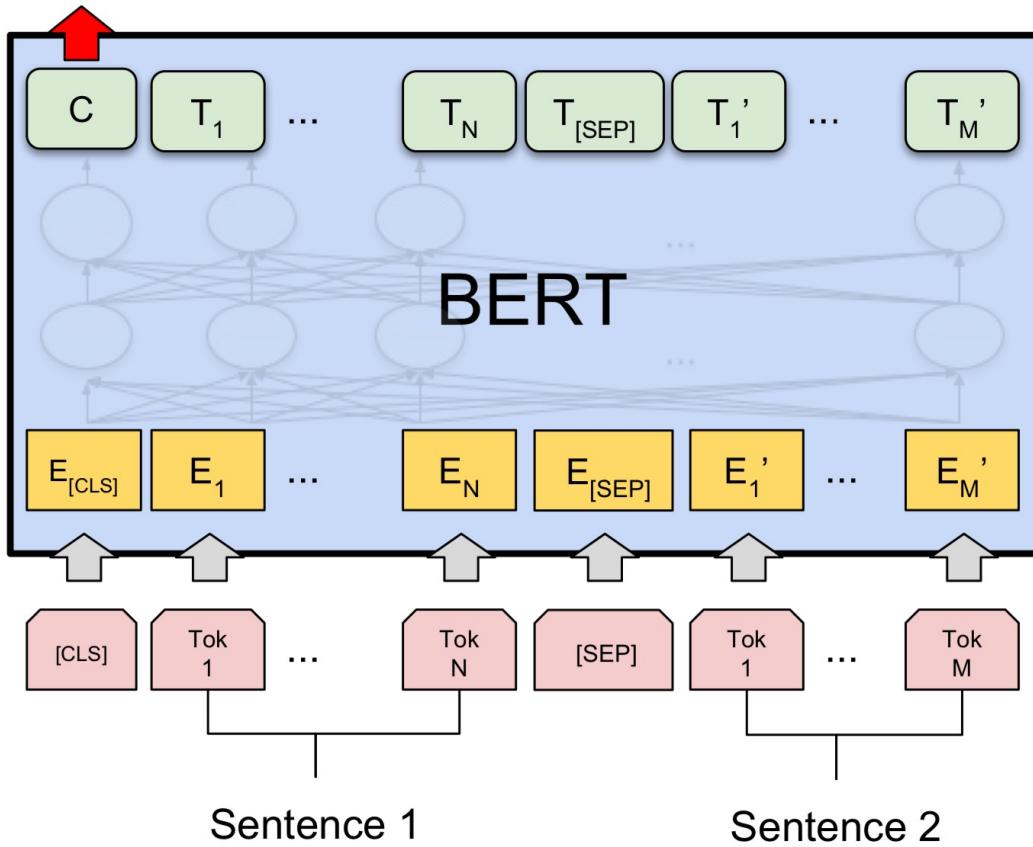
BERT — Single Sentence Classification Tasks



E.g. Sentiment Analysis
on Stanford Sentiment Treebank

BERT — Sentence Pair Classification Tasks

Class
Label



E.g. SWAG dataset:

On stage, a woman takes a seat at the piano. She

- a) sits on a bench as her sister plays with the doll.
- b) smiles with someone as the music plays.
- c) is in the crowd, watching the dancers.
- d) nervously sets her fingers on the keys.**

A girl is going across a set of monkey bars. She

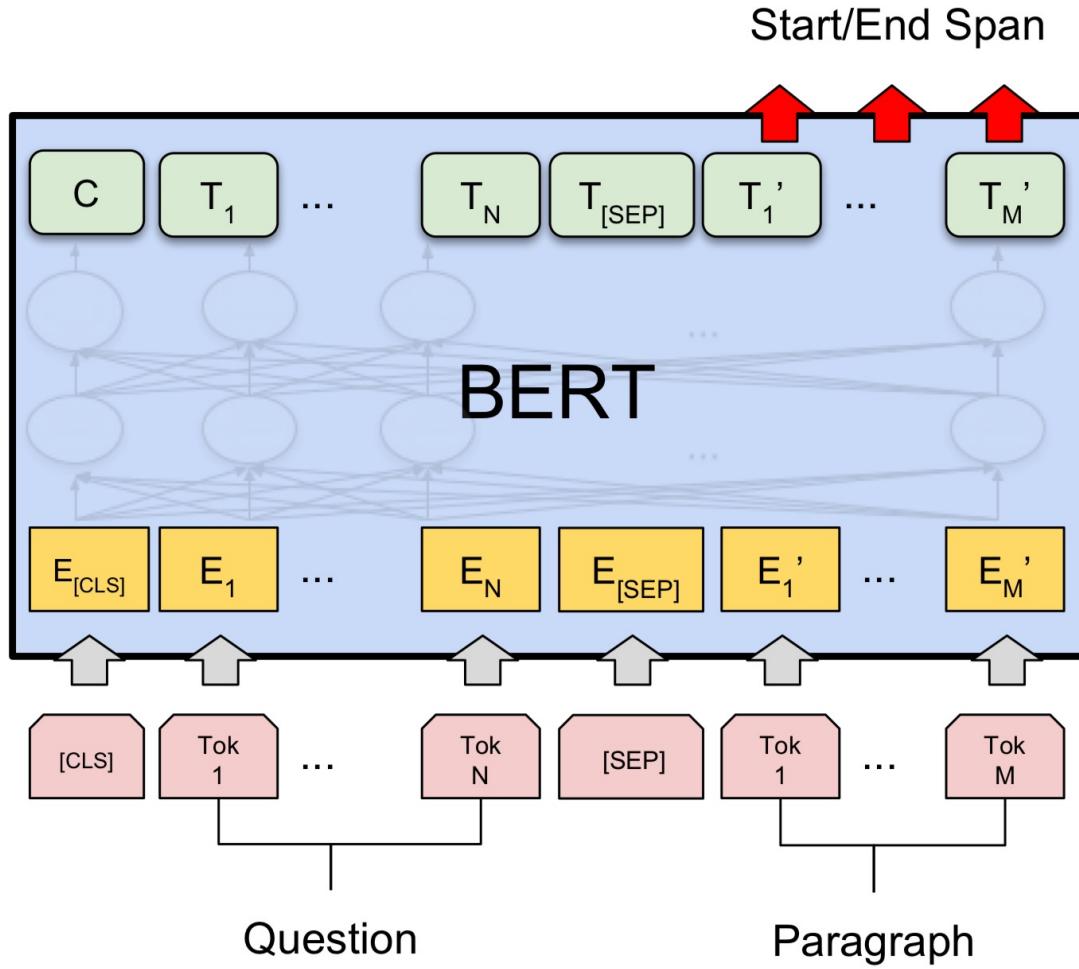
- a) jumps up across the monkey bars.
- b) struggles onto the monkey bars to grab her head.
- c) gets to the end and stands on a wooden plank.**
- d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog

- a) is placed in the kennel next to a woman's feet.**
- b) washes her face with the shampoo.
- c) walks into frame and walks towards the dog.
- d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from **SWAG**; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

BERT — Question Answering Tasks



E.g. SQuAD:

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupe****pel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

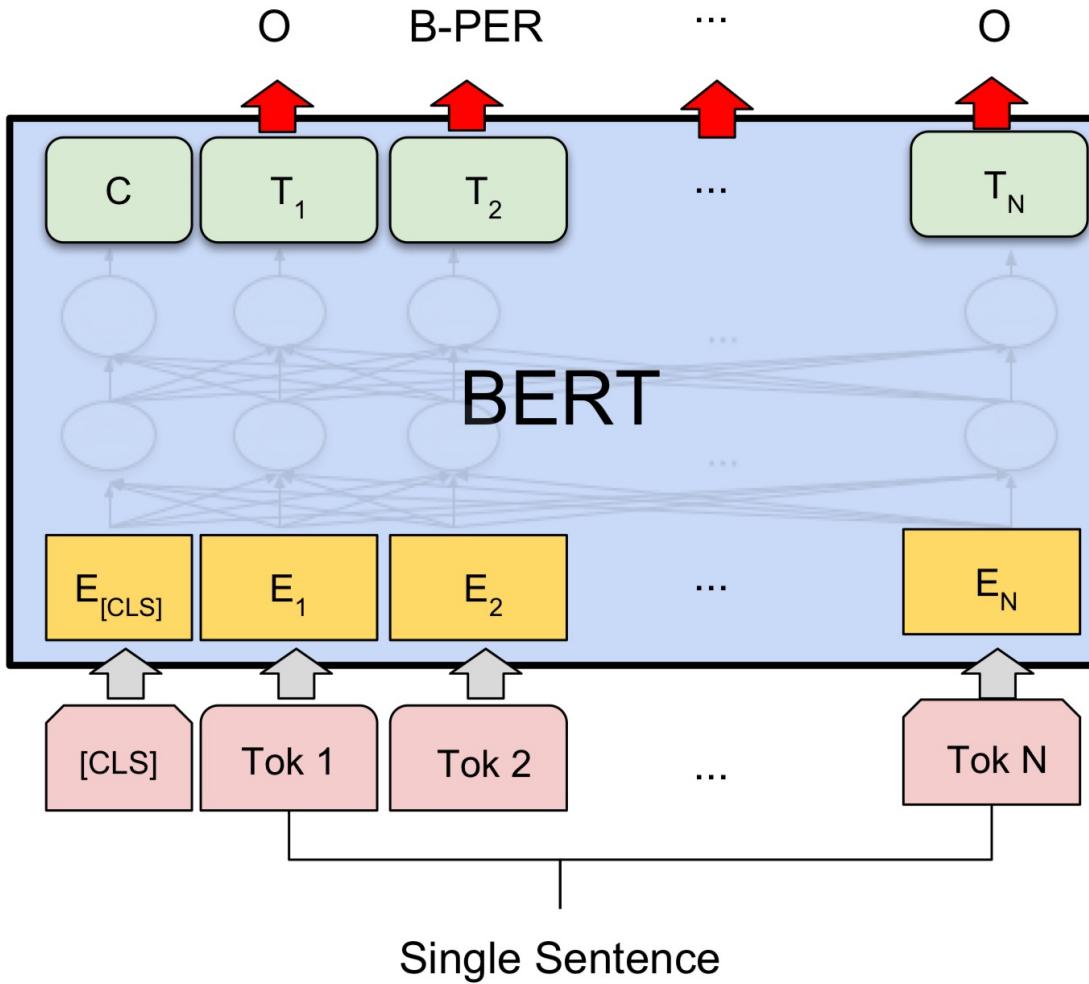
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupe**pel**

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

BERT — Single Sentence Tagging Tasks



E.g. CoNLL-2003 NER:

Named entities are phrases that contain the names of persons, organizations and locations. Example:

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].

BERT — Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Table 1: GLUE Test results, scored by the evaluation server (<https://gluebenchmark.com/leaderboard>). The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.⁸ BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

BERT — Derivatives

- Since then various derivatives have been developed up to now
 1. [BERT](#) (from Google) released with the paper [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#) by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.
 2. [RoBERTa](#) (from Facebook), released together with the paper [Robustly Optimized BERT Pretraining Approach](#) by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov.
 3. [DistilBERT](#) (from HuggingFace) released together with the paper [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#) by Victor Sanh, Lysandre Debut and Thomas Wolf. The same method has been applied to compress GPT2 into [DistilGPT2](#).
 4. [CamemBERT](#) (from FAIR, Inria, Sorbonne Université) released together with the paper [CamemBERT: a Tasty French Language Model](#) by Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djame Seddah, and Benoît Sagot.
 5. [ALBERT](#) (from Google Research), released together with the paper [ALBERT: A Lite BERT for Self-supervised Learning of Language Representations](#) by Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut.
 6. [XLM-RoBERTa](#) (from Facebook AI), released together with the paper [Unsupervised Cross-lingual Representation Learning at Scale](#) by Alexis Conneau*, Kartikay Khandelwal*, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov.
 7. [FlauBERT](#) (from CNRS) released with the paper [FlauBERT: Unsupervised Language Model Pre-training for French](#) by Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, Didier Schwab.

• • •



GPT-2

GPT-2 — Idea

- Train one **giant language model** (1.5B parameters) on a **giant text corpus** (8M web pages ~ 40GB text)!
 - Model each task as a text completion task
 - Let the model output the answer
-
- **Text summarisation:** Prompt model with „<long text>. TL;DR: “
 - **Translation:** Condition on example translations („<english sentence> = <french sentence>“), then prompt with „<english sentence> = “
 - **Reading Comprehension:** Prompt model with „<text>. Q: <question>. A: “

→ Do not fine-tune network on target task: **Zero-shot setting**

GPT-2 — Question Answering

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

GPT-2 Applications

- TabNine Code Completion: Fine-tuned on 2 million GitHub files

```
1 import os
2 import sys
3
4 # Count lines of code in the given directory, separated by file extension
5 def main(directory):
6     line_count = {}
7
8
9
10
11
12
13
14
15
16
17
18
19
```

- Write with Transformer – <https://transformer.huggingface.co>

GPT-2 Applications

Thor: The Tesseract belongs on Asgard, no human is a match for it.

Tony turns to leave, but Steve stops him.

Steve: You're [...] leaving alone!

Tony: You know I can't do that.

He turns to leave again, but Loki cuts in.

destroying the rocks impacted, said Nicholas McCarthy, director of NASA's Near-Earth object Program. "We won't start to

Loki: You see the damage from a reverberating impact sampled for

Loki cuts in again. Steve sees the words

his new enemy, but they are being chased by soldiers. As

and his enemies try to dodge the soldiers one by one, a soldier

enemies, but they are being chased by soldiers. As

Steve: Duh, I'm not going to let that will unleash

towards the Earth will be ruined.

Steve: What's the point of living if you're just going to be middle-aged, shaved canines.

They are running away from their enemies, but they are being chased by soldiers. As

Steve and his friends try to dodge the soldiers one by one, a soldier

them. Steve throws up a spear at the soldier.

Loki: You're nothing like Thor!

GPT-2 — Ethical Considerations

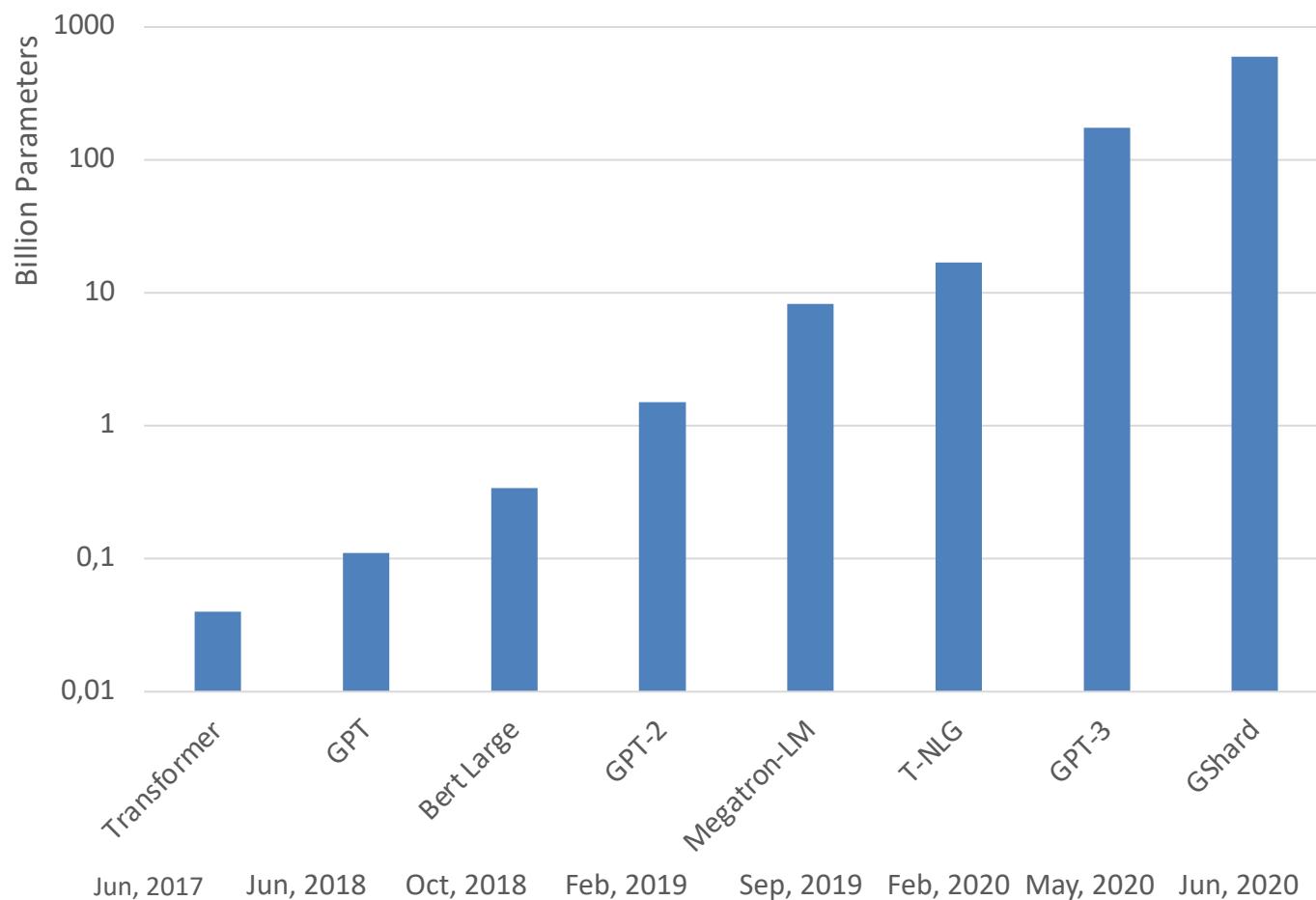
- The largest model has 1.5B parameters
- That is so 2019...
- Open source
- On GitHub
- well...
—
- Biggest concern: “The parameters, were only shared with researchers, who are working to improve societal preparedness for large language models”

1.5 billion parameters?

That is so 2019...

dia

Language Model Sizes



GPT-3 — Example

Describe a layout.

Just describe any layout you want, and it'll try to render below!

A div that contains 3 buttons each with a random color.

Generate





Evaluating Language Models — Perplexity

- How do we know if our models are any good?
 - And in particular, how do we know if one model is better than another?
- Generated texts from models with more parameters look better
- That is, they sound more like the text the model was trained on
- Can we make that notion operational?
- One such metric is perplexity → Intrinsic evaluation

Evaluating Language Models — Perplexity

- Perplexity is the (inverse) probability of the test set (assigned by the language model), normalized by the number of words:

$$PP(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}}$$

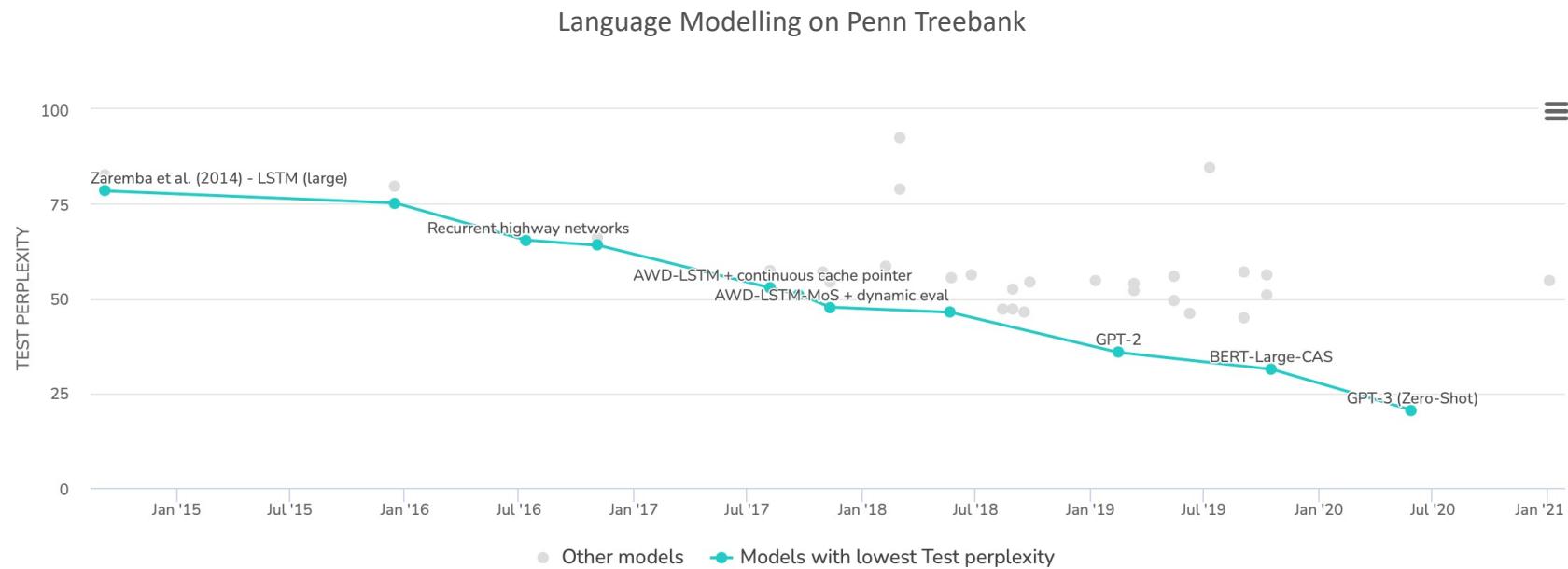
$$= \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}}$$

- Approximation:

$$PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_1, \dots, w_{i-1})}}$$

Evaluating Language Models — Perplexity

- Lower perplexity means a better model
- Minimizing perplexity is the same as maximizing probability
- The best language model is one that best predicts unseen text



<https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word>

Language Models are Few-Shot Learners

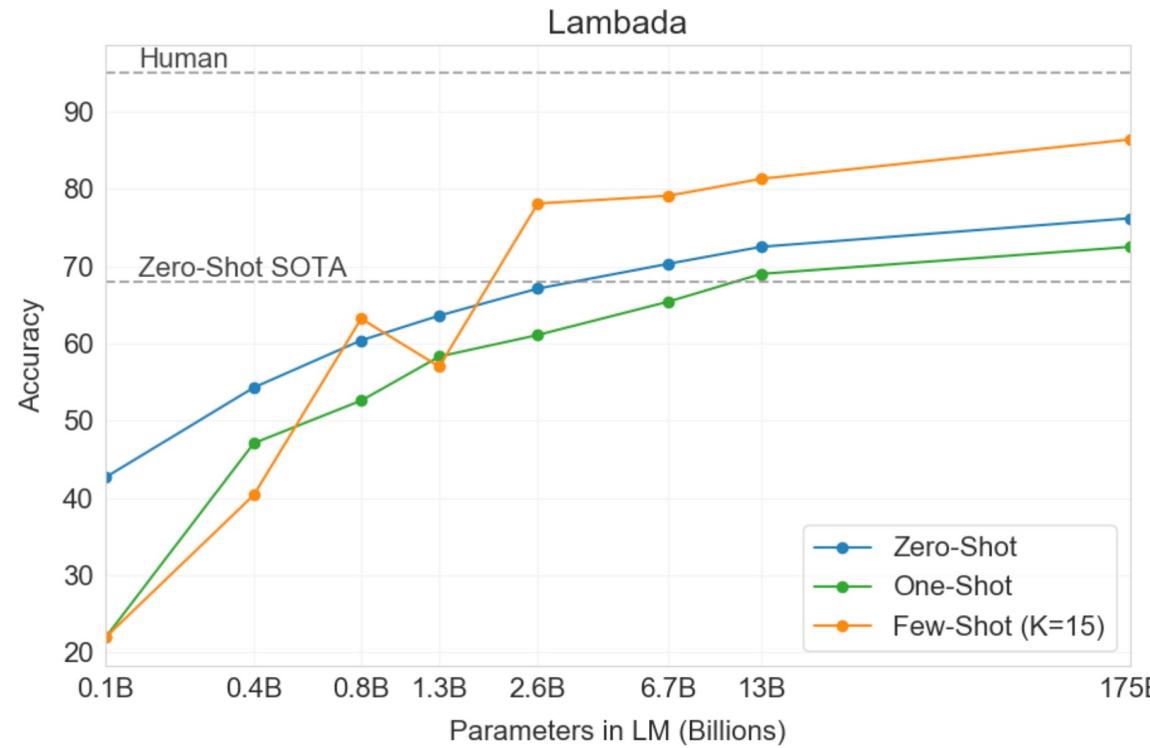


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text.

Language Models are Few-Shot Learners

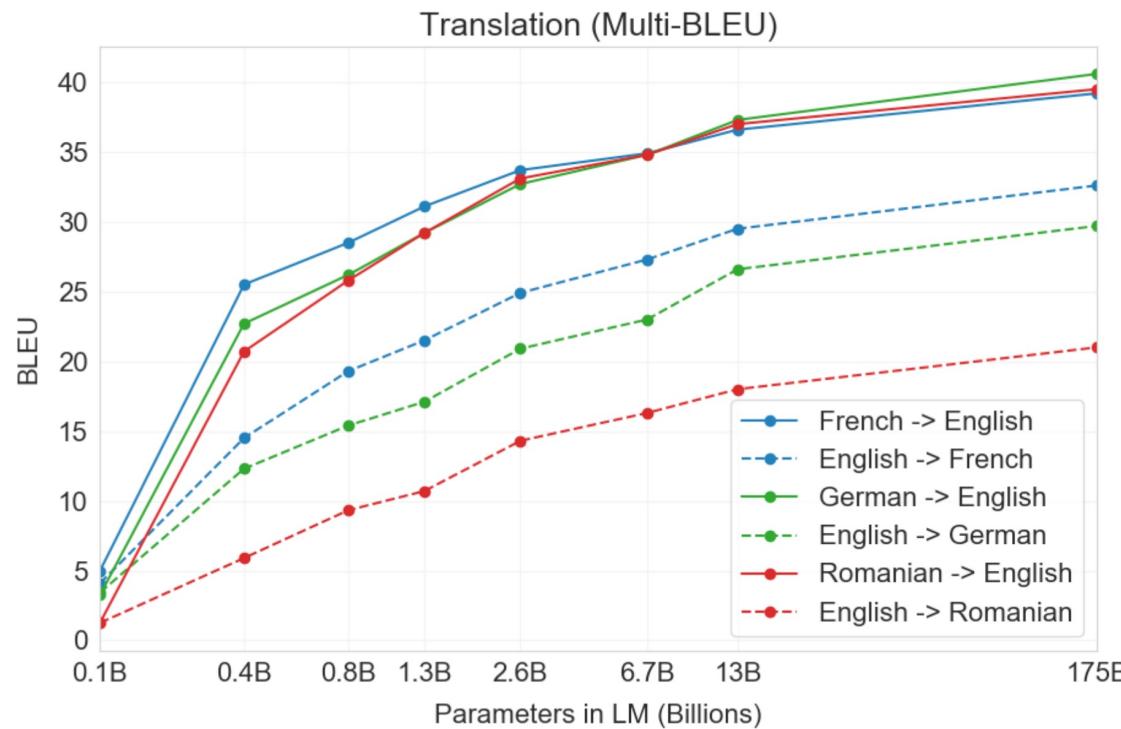


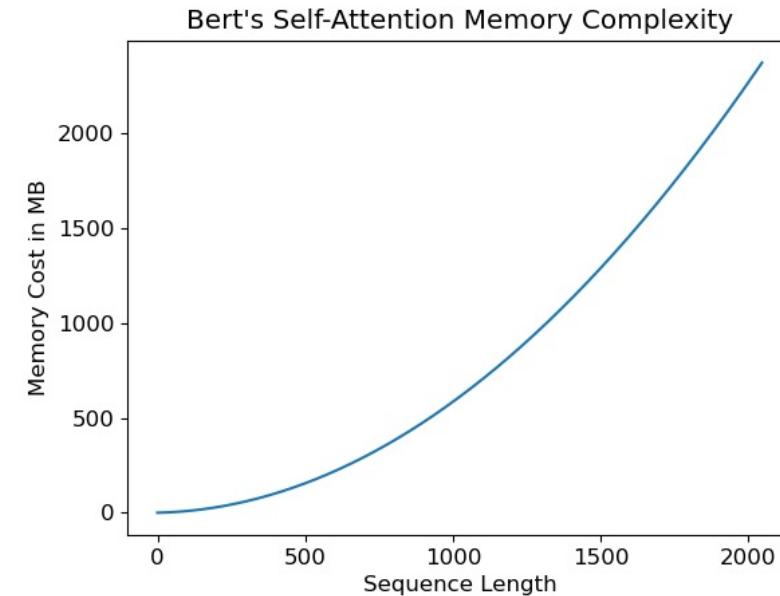
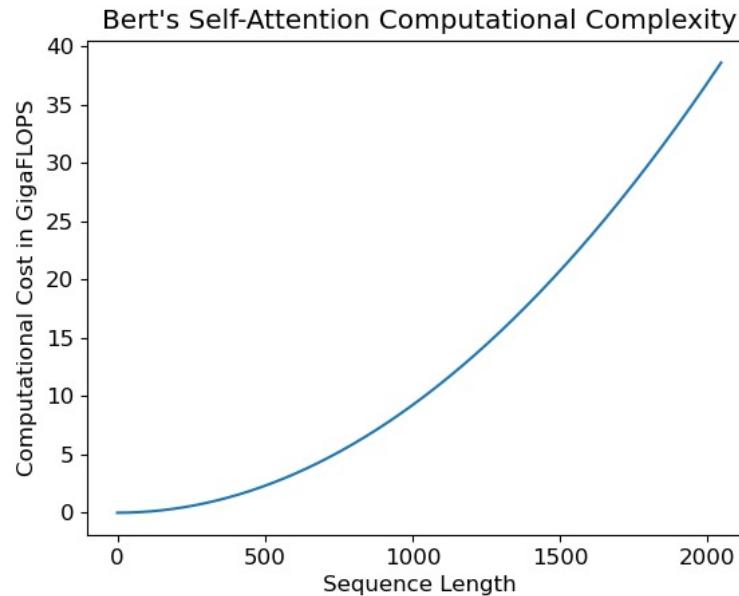
Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be stronger than translation from English.

Problems of Big Language Models

- Let's look at BERT's parameters
 - n : Sequence length of 512
 - h : 12 self-attention heads
 - d : Hidden size of 768, 12 heads $\rightarrow 64$ ($768/12$)
- Let's look at BERT's complexity per layer
 - Computational: $O(hdn^2)$
 - Memory: $O(hdn + hn^2)$
- Is this good? Let's look at the total required memory
 - $12 \times 64 \times 512 + 12 \times 512 \times 512 = 393.216 + 3.145.728 = 3.538.944$ floats
 - $3.538.944 \times 12$ Encoder-Layer = 42.467.328 floats
 - This is 170 MB of memory required for a single sample!
- What is the key factor here?

Problems of Big Language Models

- Memory and computational complexity both scale to the square of the sequence length



- Many improvements have been made since, let's look at three

1. Q8BERT — Quantized 8Bit BERT

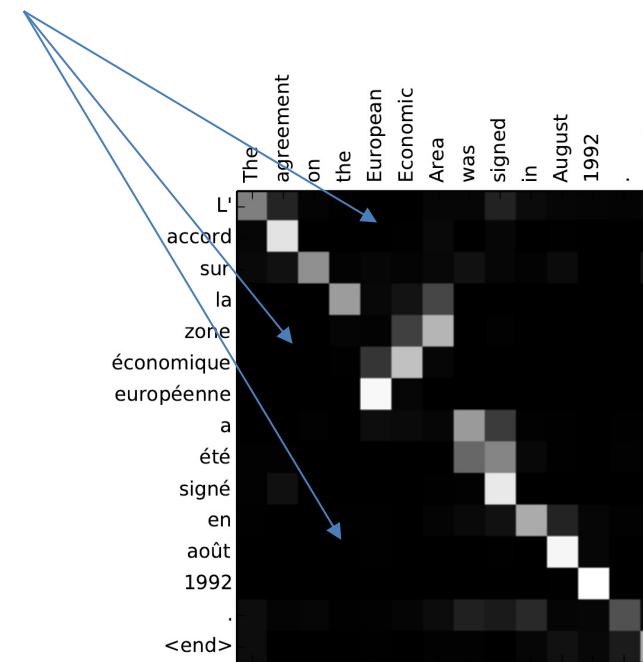
- Floating point numbers use 4 bytes of memory
- Can we get away with using only 1 byte?
- Yes! Zafrir et. al use linear quantization during inference:

$$\begin{bmatrix} [-0.52708159 \ 0.58938589 \ -0.99298076] \\ [-0.80074533 \ -0.24025461 \ -0.70421488] \\ [0.54495217 \ -0.23490439 \ 0.76274836] \end{bmatrix} \xrightarrow{\hspace{1cm}} \begin{bmatrix} [-68 \ 75 \ -127] \\ [-103 \ -31 \ -91] \\ [69 \ -31 \ 97] \end{bmatrix}$$

- The model needs to learn to bridge the “quantization error gap”
- Therefore the rounding effect is simulated during training
- Inference not only requires considerably less memory (~25%), but also runs faster on optimized hardware (~4x)

2. Reformer – The Efficient Transformer

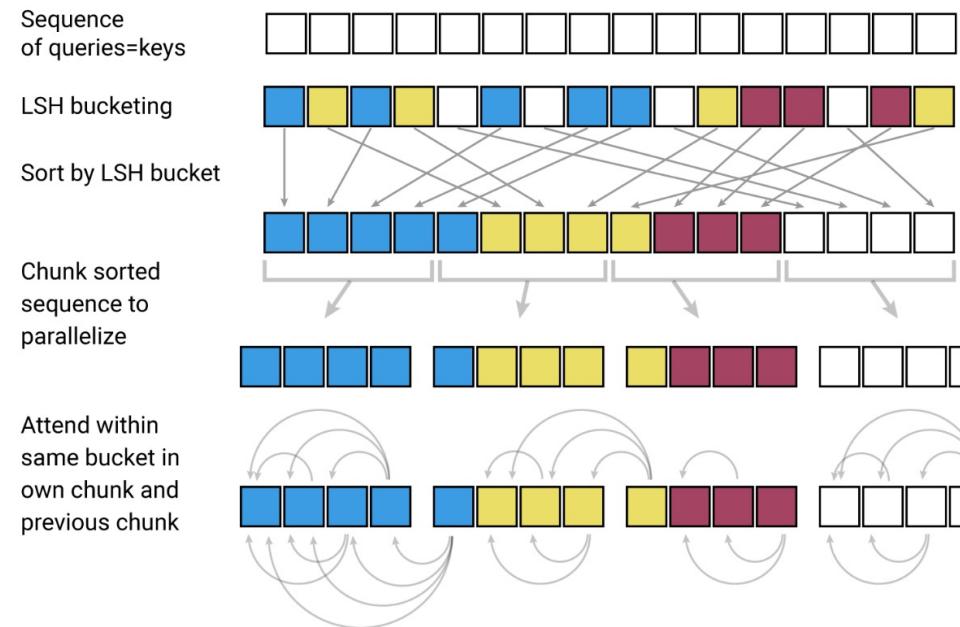
- Remember how a compatibility function is used to calculate attention between queries and keys
 - In the simplest case this is the dot product, i.e. vector similarity
 - Since incompatible vectors contribute little to the sum, we can ignore them



Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

2. Reformer – The Efficient Transformer

- Kitaev et. al introduce “Locality Sensitive Hashing (LSH)“
 - Divide vectors into different “buckets” according to similarity
 - Attention now only attends to vectors of the same bucket



- Thus, the computational and memory complexity reduces to $O(L \log L)$

3. DistilBERT — A Distilled Version of BERT

- Of course the sheer number of parameters is also problematic
 - Magnitudes of gigabytes are infeasible for IoT and mobile devices
- **Knowledge Distillation:** Train a smaller model (student) to reproduce the behavior of a larger model (teacher)
- Sanh et. al reduce the size of BERT by 40%, make it 60% faster, and still retain 97% of its capabilities

Conclusion

- Recent trend in NLP: **Transfer Learning**
 - Train a model on a task with a big amount of training data
 - Then modify the model and fine-tune on the target task
 - Many improvements on a day to day basis, e.g. Linformer (Jun, 2020)
- Word embeddings come from transfer learning
 - They are trained on large corpora
 - Then they are used in other tasks (with or without fine-tuning)
- Recent developments: Pretrain a language model and directly adapt it to the target task → No word embeddings
- Word embeddings may go extinct if this trend continues 😔