Attention & Transformers for NLP Tasks

By: John Cisler and Ben Paulson

BIG QUESTION

Why are transformers, and attention, used in some of the most effective deep-learning models today for NLP?

“Attention Is All You Need” & Language Translation

Paper Key Takeaways

Attention Is All You Need

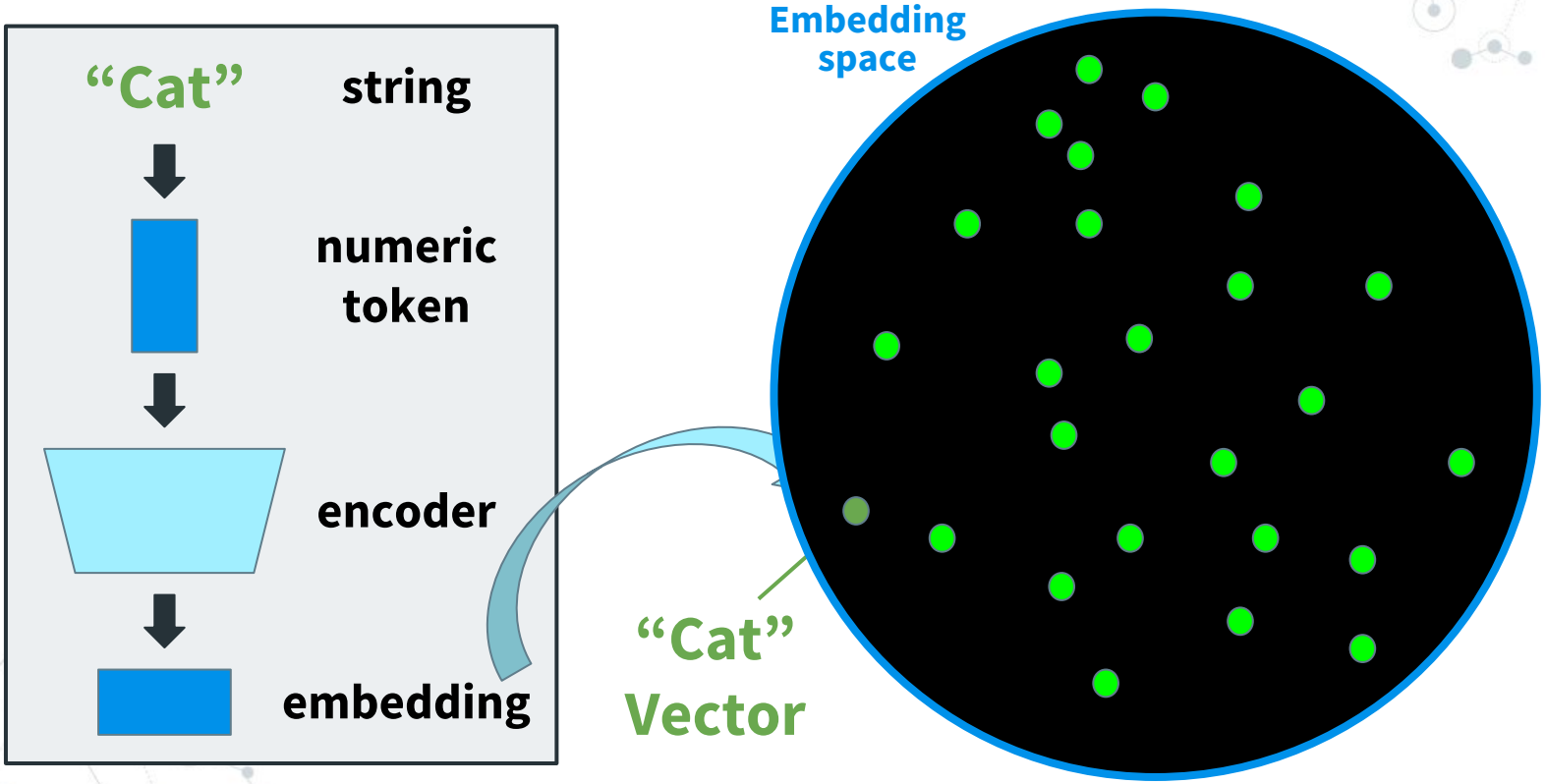
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

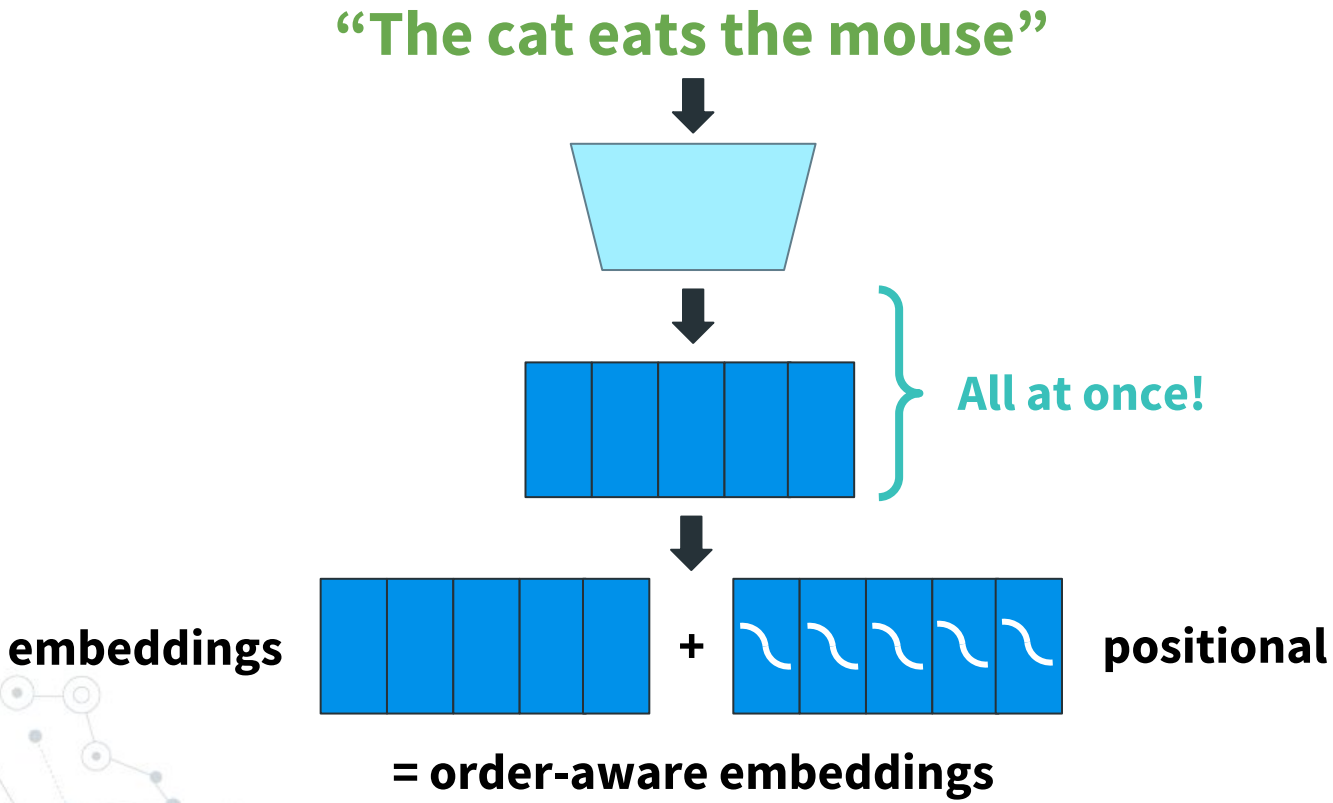
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

- Language Translation
  - English to French/German
- Transformer
  - Seq2Seq Model
  - High Throughput
  - Easily parallelizable
  - Mitigated long-range dependency issue
  - State-of-the-art results

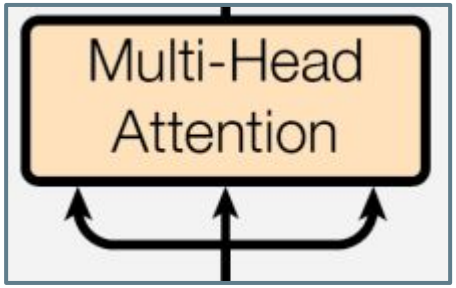
High Throughput – Embeddings



High Throughput – Embeddings

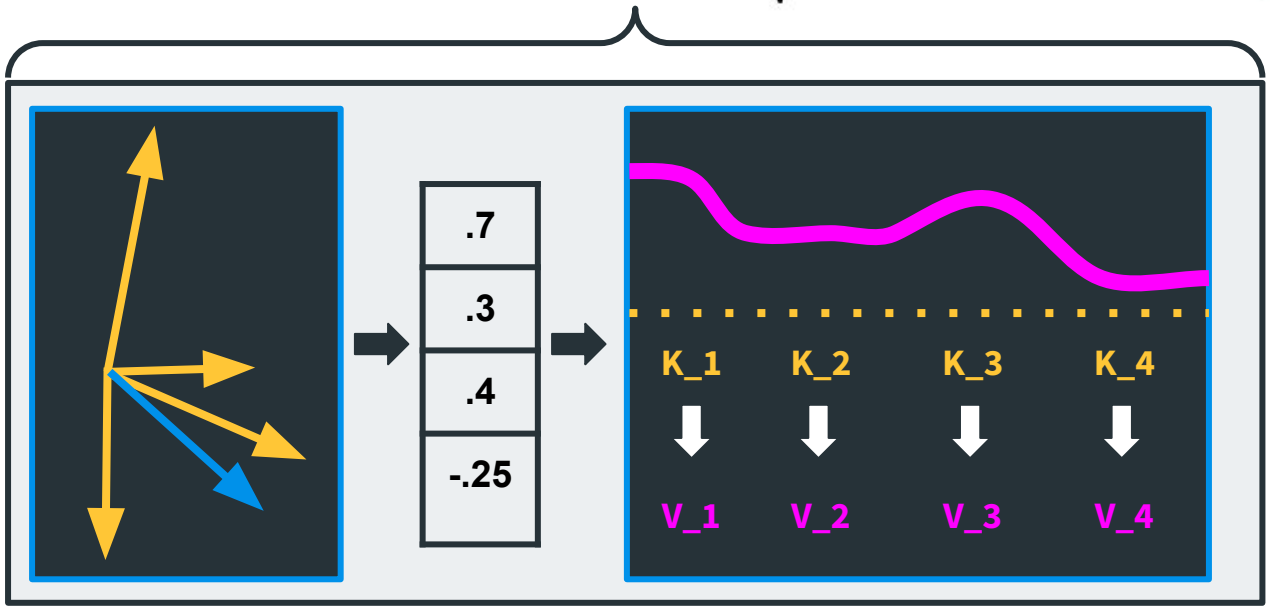


Long Range – Attention



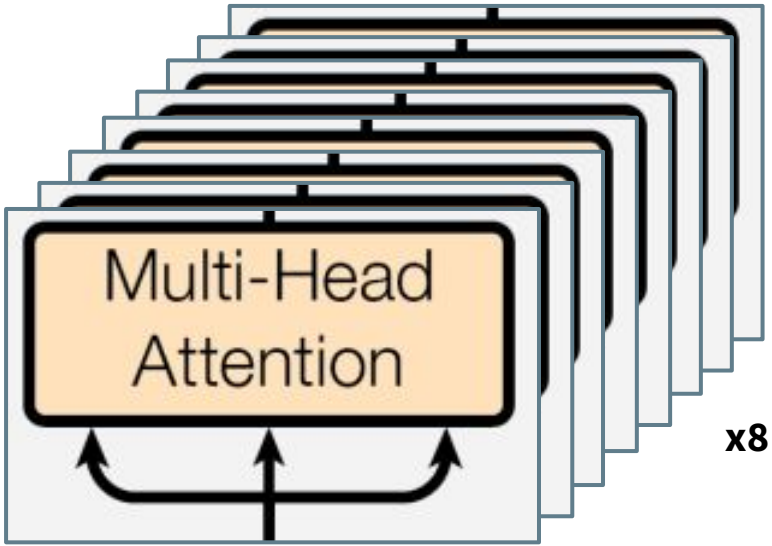
A(Q, K, V)

➡ $A(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) * V$



Parallelization – Multi-Headed

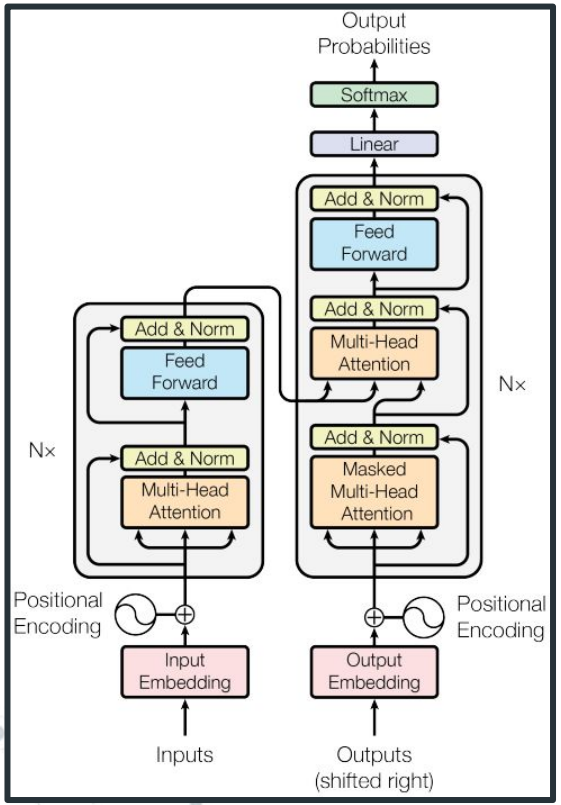
$A(Q, K, V) =$



x8 “Heads”!

Multiple heads capture more abstract features

All Created the Transformer!

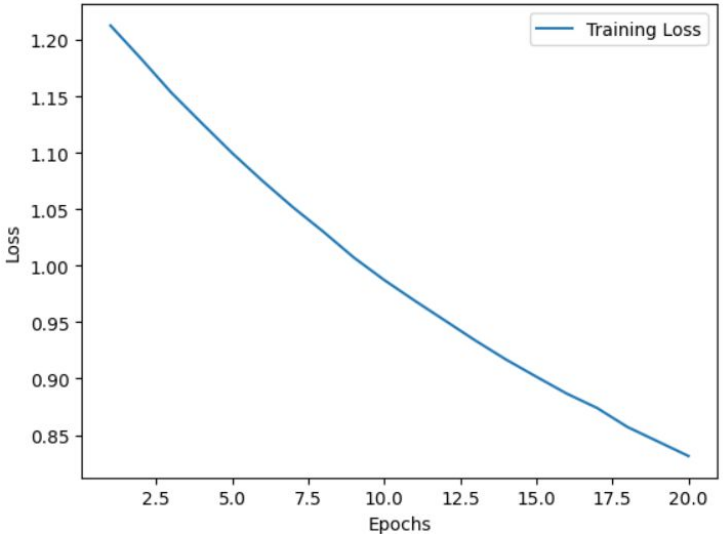


- High Throughput
- Captures long-range dependencies
- Parallelizable

Next time, we'll walk through inputs as they travel through this transformer

Language Translation

```
Epoch 10/20
810/810 [==============================] - 181s 224ms/step - loss: 0.9873 - masked_accuracy: 0.7576 - val_loss: 2.1494 -
val_masked_accuracy: 0.6296
Epoch 11/20
810/810 [==============================] - 182s 225ms/step - loss: 0.9690 - masked_accuracy: 0.7608 - val_loss: 2.1680 -
val_masked_accuracy: 0.6274
Epoch 12/20
810/810 [==============================] - 180s 223ms/step - loss: 0.9513 - masked_accuracy: 0.7642 - val_loss: 2.1657 -
val_masked_accuracy: 0.6332
Epoch 13/20
810/810 [==============================] - 182s 224ms/step - loss: 0.9336 - masked_accuracy: 0.7670 - val_loss: 2.1649 -
val_masked_accuracy: 0.6321
Epoch 14/20
810/810 [==============================] - 180s 222ms/step - loss: 0.9169 - masked_accuracy: 0.7703 - val_loss: 2.1904 -
val_masked_accuracy: 0.6289
Epoch 15/20
516/810 [==================>...........] - ETA: 1:04 - loss: 0.8831 - masked_accuracy: 0.7778
```




```
sentence = 'este é um problema que temos que resolver.'
ground_truth = 'this is a problem we have to solve .'

translated_text, translated_tokens, attention_weights = translator(
    tf.constant(sentence))
print_translation(sentence, translated_text, ground_truth)
```

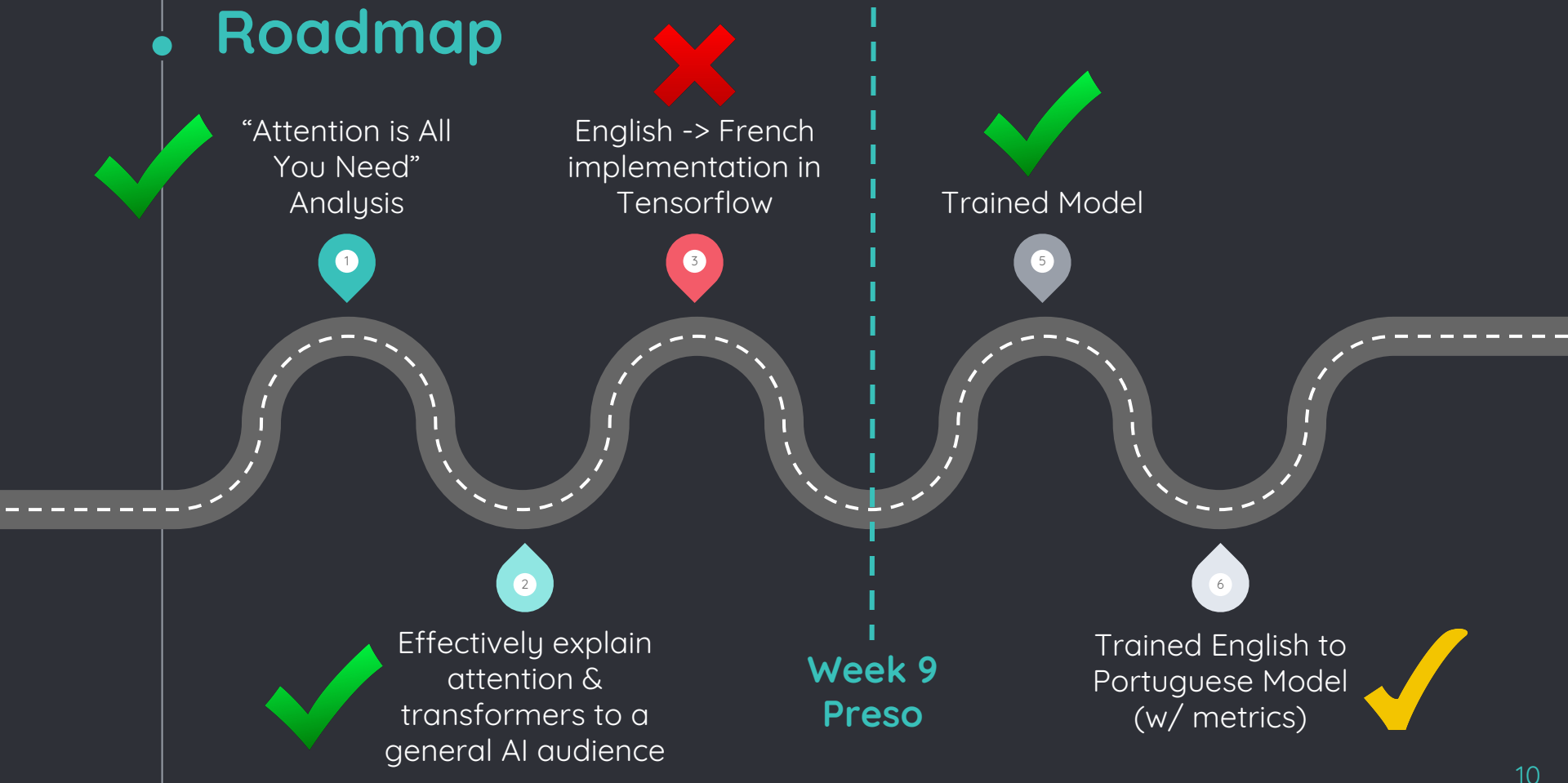
```
Input:         : este é um problema que temos que resolver.
Prediction     : this is a problem that we have to solve .
Ground truth   : this is a problem we have to solve .
```

4 Layers vs 6 Layers

30 min vs 20 hours

BLEU Score

Roadmap


“Attention is All You Need” Analysis
1
Effectively explain attention & transformers to a general AI audience
2
English -> French implementation in Tensorflow
3
Week 9 Preso
Trained Model
5
Trained English to Portuguese Model (w/ metrics)
6