Advanced neural network

Simen Dymbe og Torstein Nordgård-Hansen

Outline

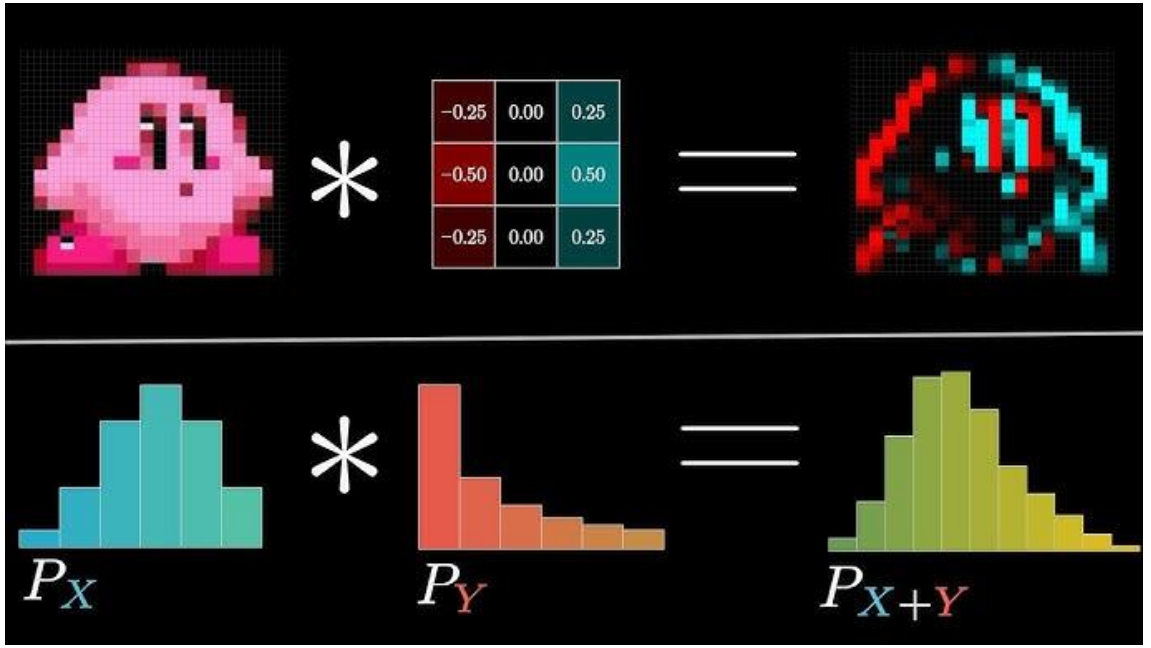
- Advanced neural networks
- Convolutional neural networks
- Transformer networks

Why advanced architectures

- Adding more neurons -> overfitting
- Training sets are usually finite
- Backpropagation will not fix architectures
- Previous states grows fast
- Today:
  - Convolutional
  - Transformer
- For you:
  - (old) cheat sheet: https://www.asimovinstitute.org/neural-network-zoo/

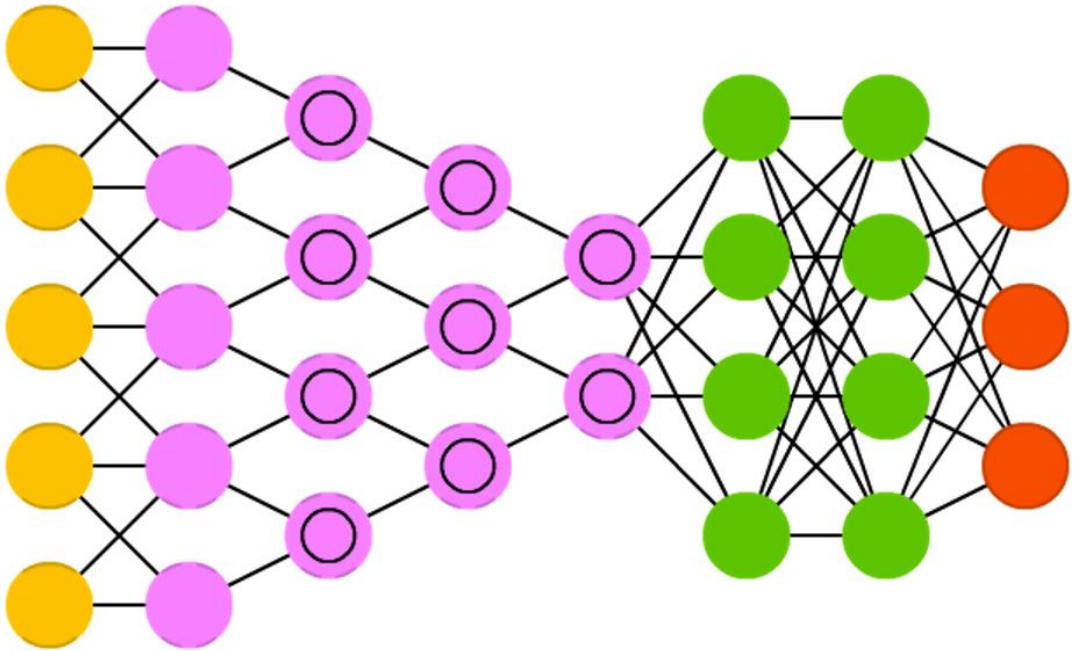
Convolutional neural networks

- Ideal for local features
- Primarily image and audio uses
- Integrated as layers in FFNNs
- Backpropagation finds kernel



Design choices and hyper parameters

- If combined with FFNNs, where and how?
- How wide and deep kernels?
- How many input nodes?



Cheatsheets

Transformers

Quick introduction

- Transformers introduced in 2017
- Cited 140 833 times
- Applications in:
  - Text processing
  - Speech processing
  - Image processing
  - etc

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

Words (or other things) as vectors

way up here

Word example

mansion

house

domicile

apartment

caravan

vehicle

bicycle

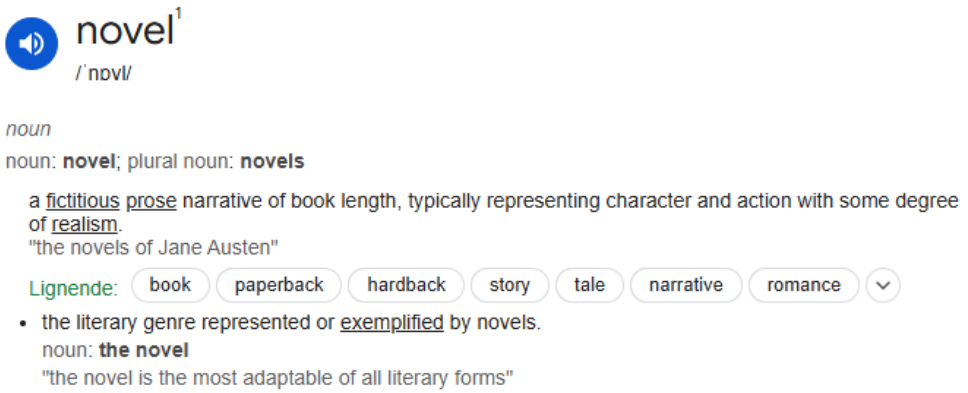
car

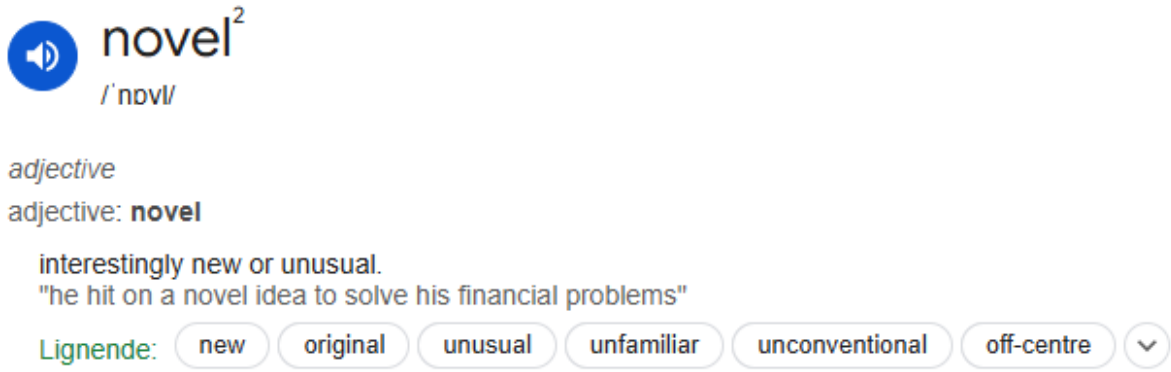
Image example

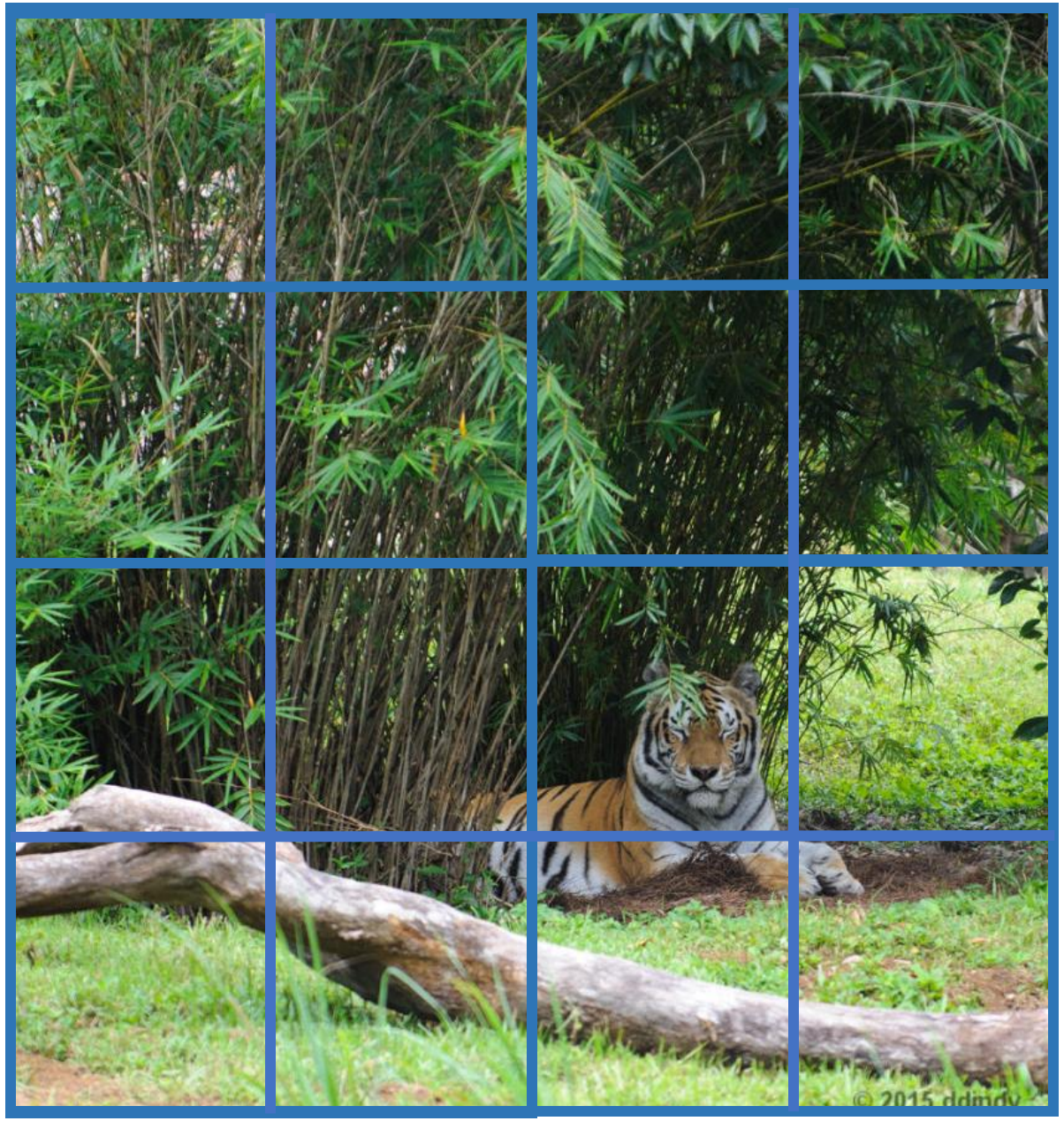
Transformers - Motivation

Reaction to the novel varied widely upon publication. Despite the number of copies sold and its widespread use in education, literary analysis of it is sparse.

- From the Wikipedia on «To Kill a Mockingbird»

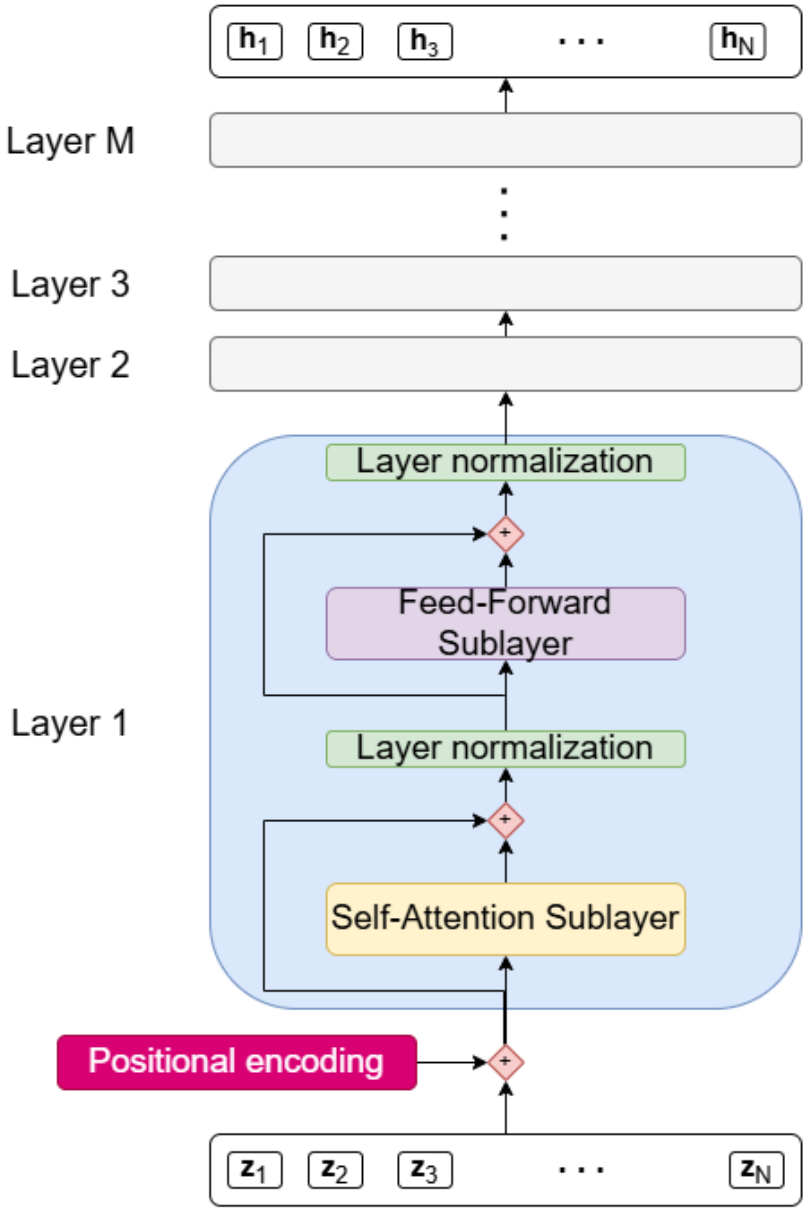





© 2015 ddmdv

The architecture

- Transformer Blocks placed in series
  - GPT-3 has 96 blocks
  - GPT-4 has 120 blocks
  - Whisper (speech) has 64 blocks

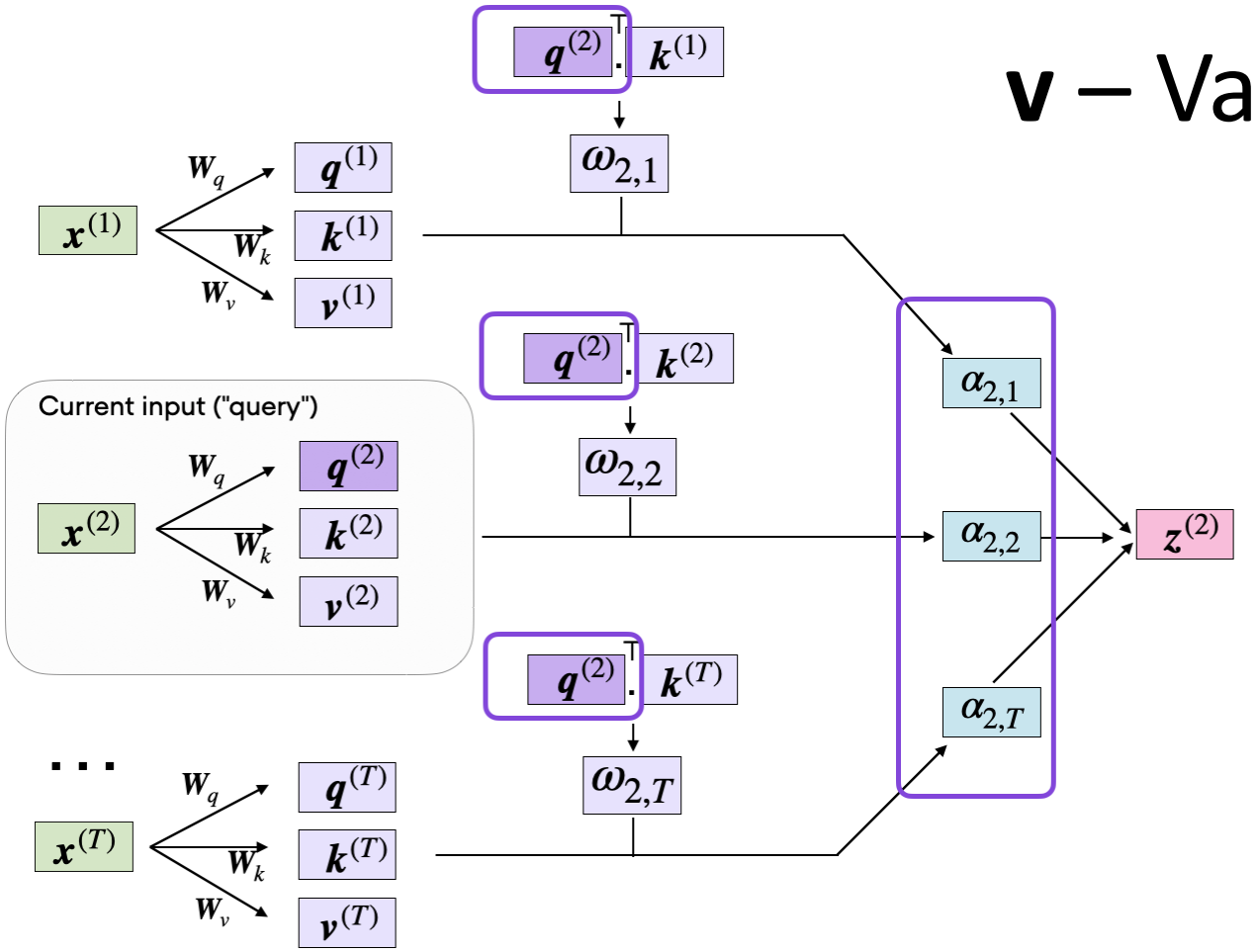


Self-attention

q – Query vector

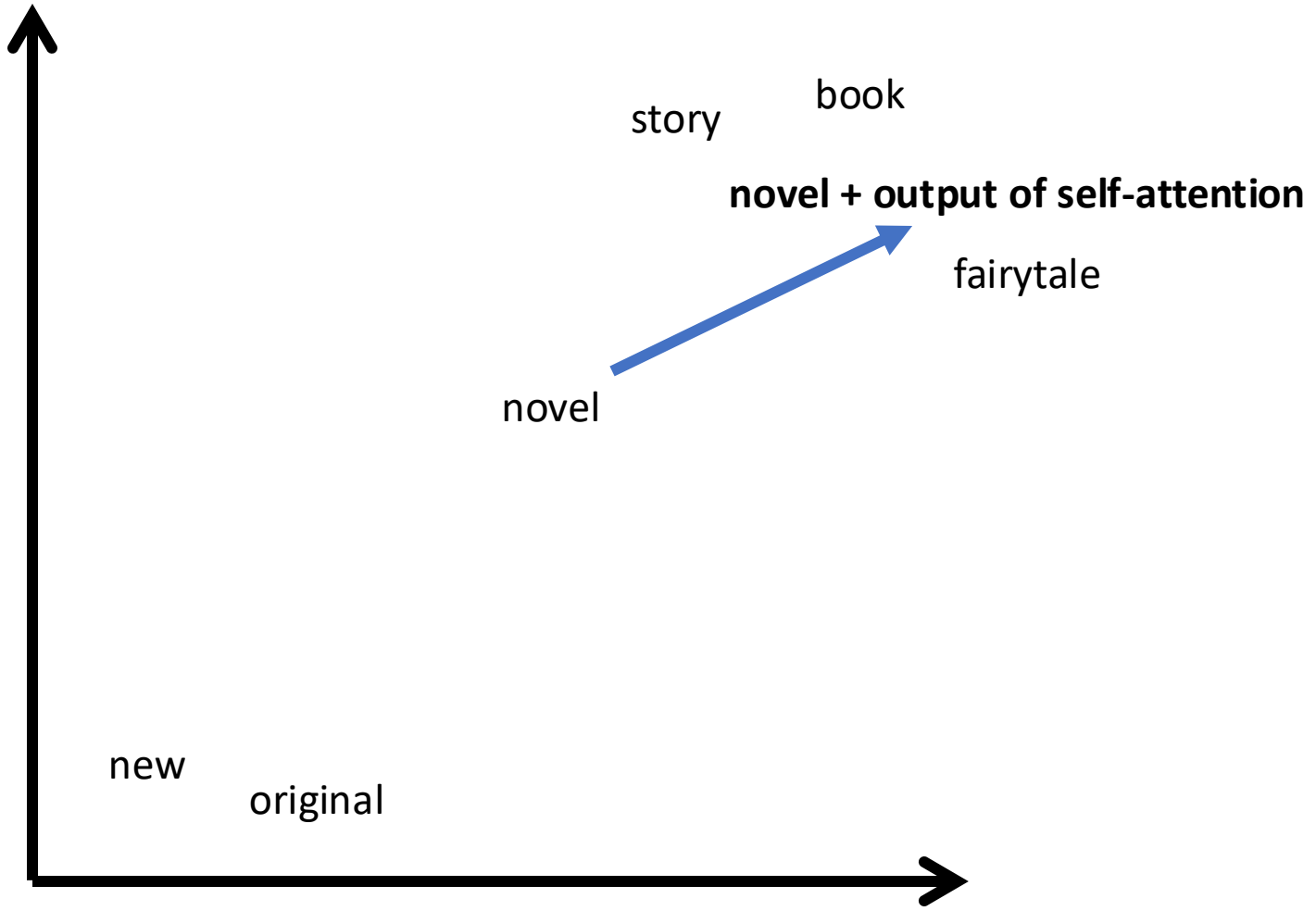
k – Key vector

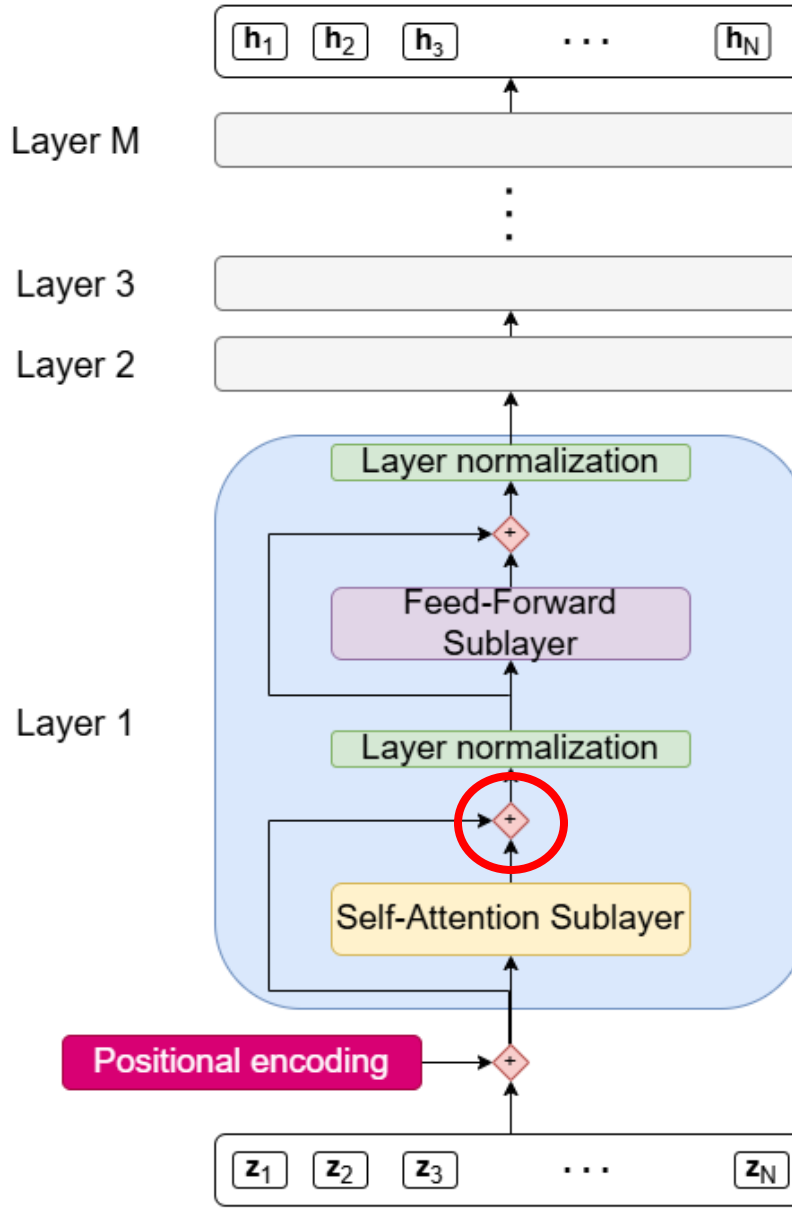
v – Value vector



where $z^{(2)} = \sum_{j=1}^{T} \alpha_{2,j} v^{(j)}$

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html


story
book
novel + output of self-attention
fairytale
novel
new
original


h1
h2
h3
hN
Layer M
Layer 3
Layer 2
Layer 1
Layer normalization
Feed-Forward Sublayer
Layer normalization
Self-Attention Sublayer
Positional encoding
z1
z2
z3
zN

Attention Visualizations

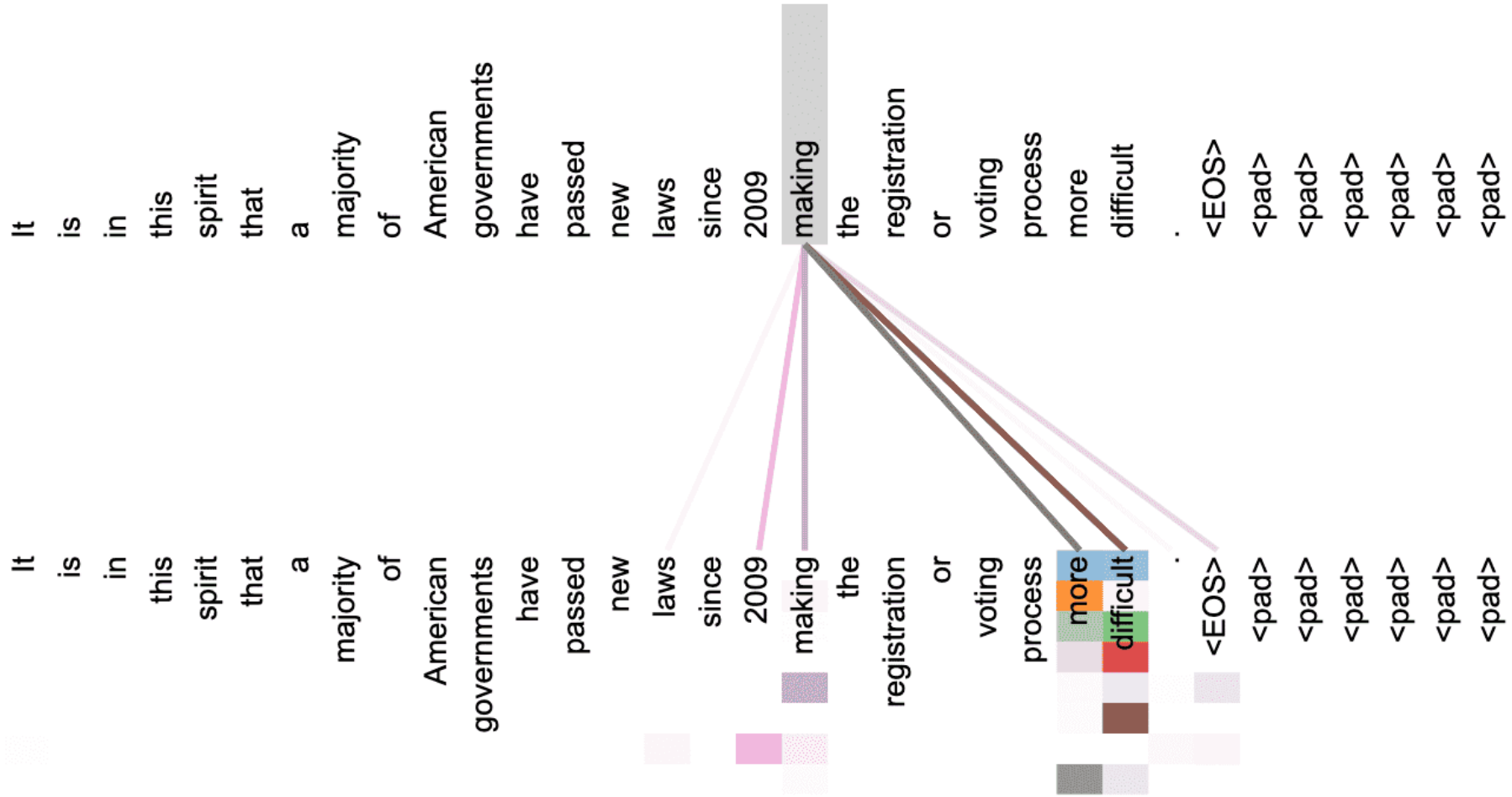


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

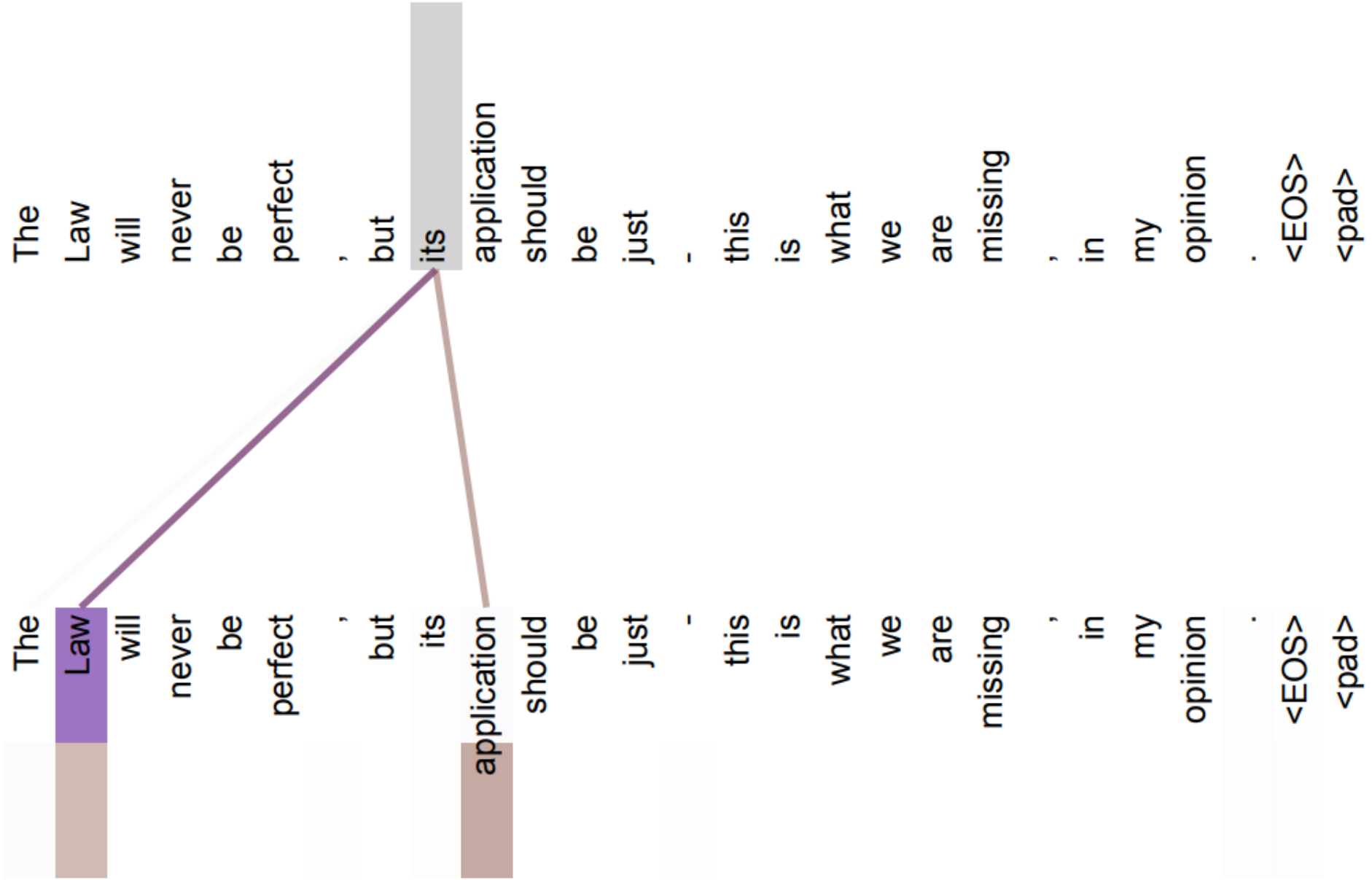


Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

Why Transformers and not CNN/RNN?

CNNs

- Great at local patterns, bad at long-range dependencies
- Easily parallelized

RNNs

- Limited handling of long-range dependencies (bottleneck problem)
  - LSTMs are better, but does not solve the problem entirely
- Hard to parallelize

Transformers

- Handles long-range dependencies
- Easily parallelized

Why not Transformers?

- Computationally expensive
- Large memory requirement

Summarize

- Processes sequences of vectors
- Works based on the self-attention mechanism
- Why use them
  - Scale well
  - Handles long-range dependencies
  - Great for text, speech, images, etc.