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

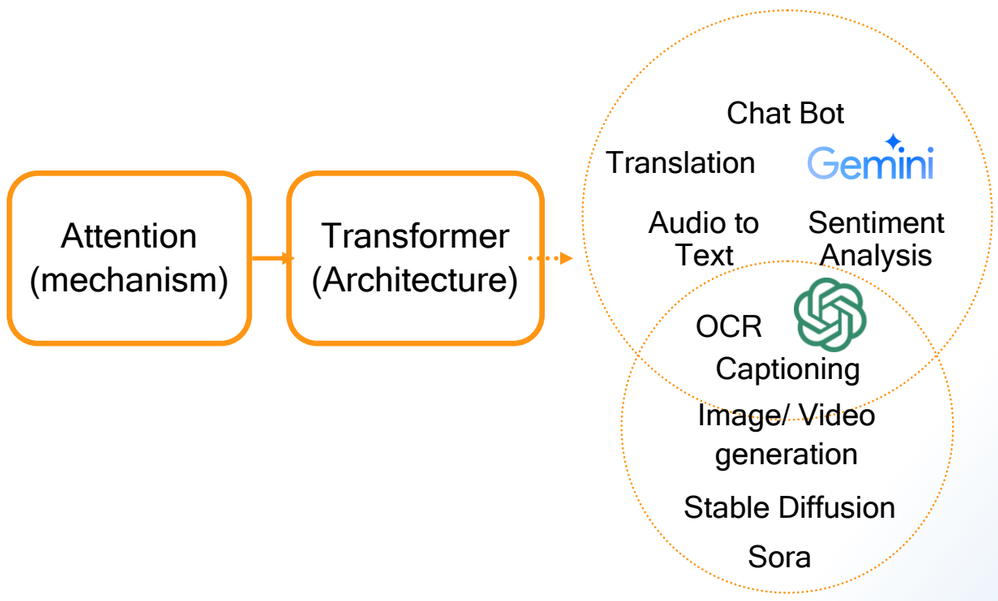
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

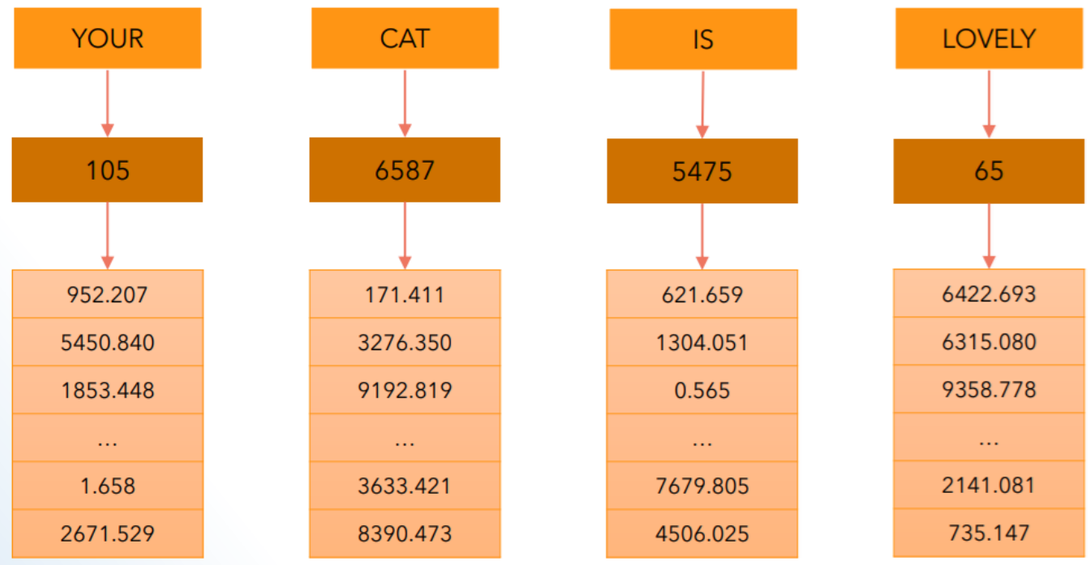
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Motivation



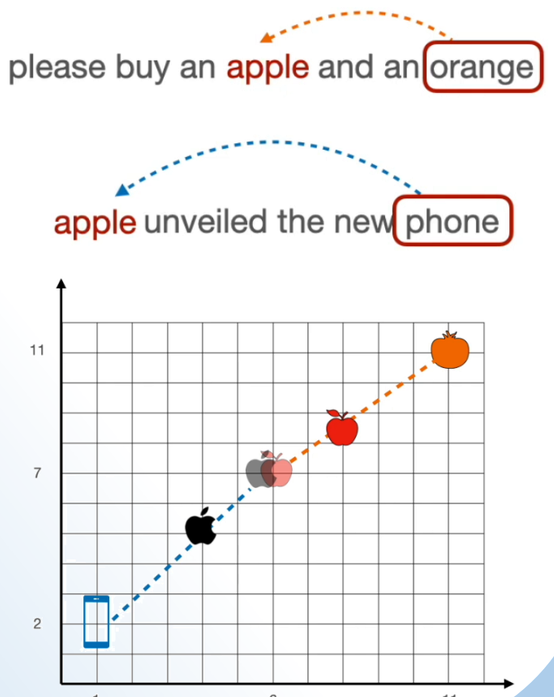
Attention	LSTM/CNN
Single step training	Training scales with length of sequence (LSTM)
Both short and Long range interaction for context	Only local interaction (CNN)

Embedding



Positional Encoding

Generally, after the embedding, the positional information is added (literally) to the embedding vector. It is independent of embedding and fixed by the position of a word in the sequence.

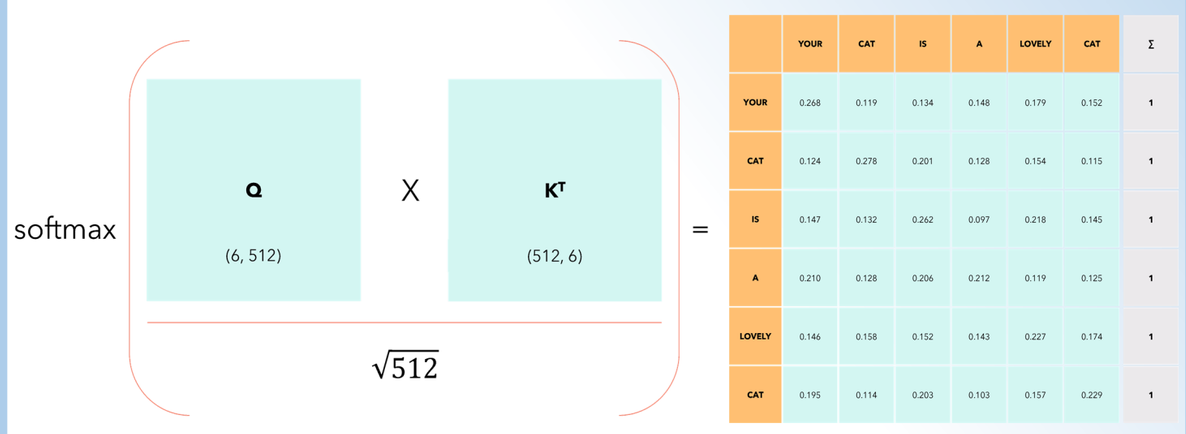


ATTENTION

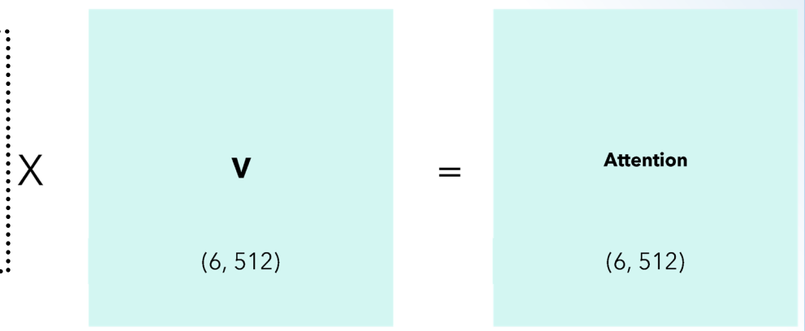
Is matrix algebra to map one sequence to another based on context.

Self Attention

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



Say there are 6 words with an embedding of size 512, then the Q, K and V matrices are of size 6x512 each.

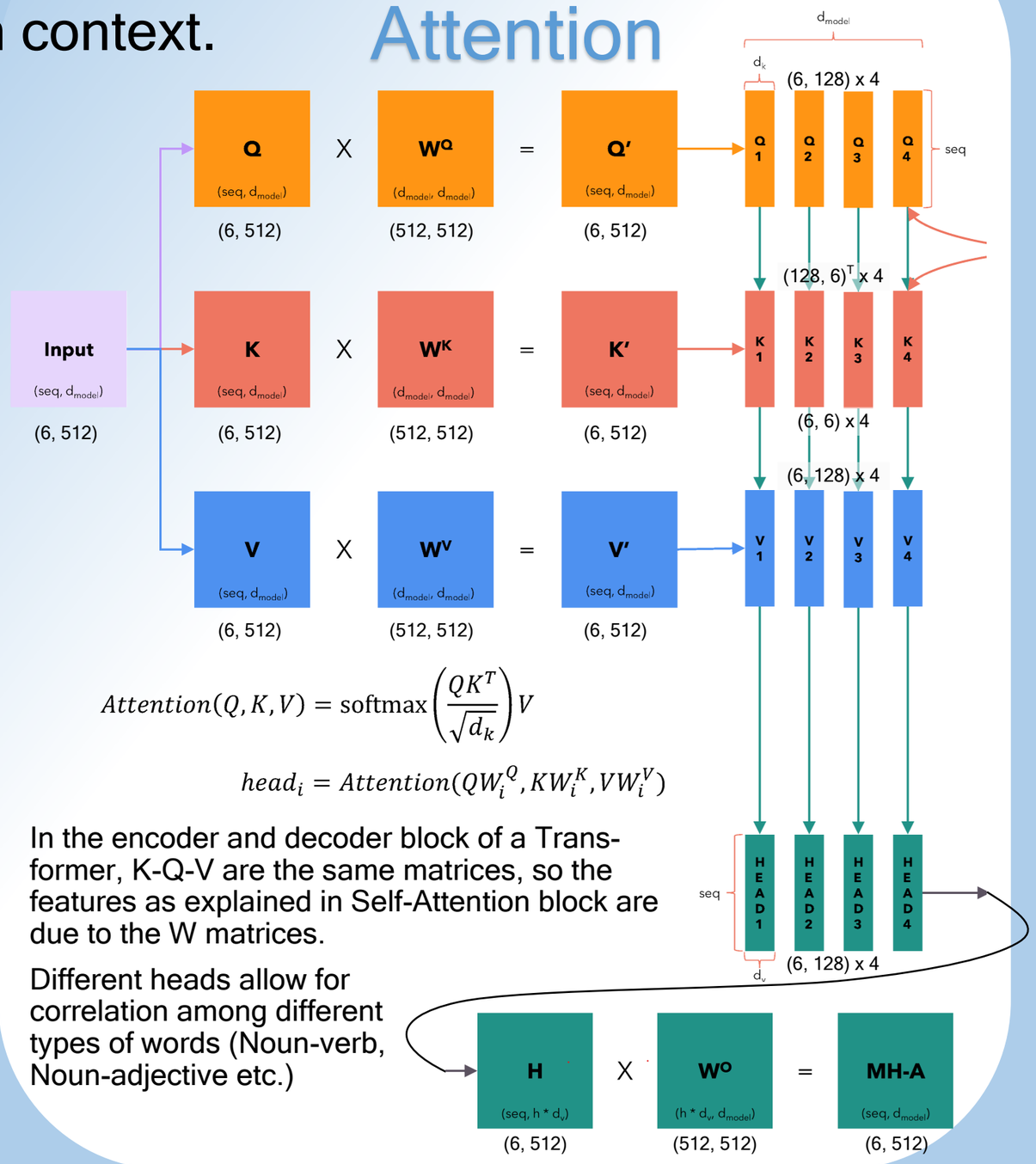


Q : (Input) Embedding in a space where related words are closer and unrelated words farther.

K : Relates the input to the output

V : (Output) Embedding in a space where words that come next in a sentence are closer (used during generation/ inference)

Multi-head Attention



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

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

In the encoder and decoder block of a Transformer, K-Q-V are the same matrices, so the features as explained in Self-Attention block are due to the W matrices.

Different heads allow for correlation among different types of words (Noun-verb, Noun-adjective etc.)

School of Computer Sciences
ରାଷ୍ଟ୍ରୀୟ ବିଜ୍ଞାନ ଶିକ୍ଷା ଏବଂ ଗବେଷଣା ପ୍ରତିଷ୍ଠାନ
राष्ट्रीय विज्ञान शिक्षा एवंअनुसंधान संस्थान
NATIONAL INSTITUTE OF SCIENCE EDUCATION AND RESEARCH

CS460 Machine Learning Posters- 2024
Pritipriya Dasbehera
pritipriya.dasbehera@niser.ac.in

References:
YouTube/@umarjamilai
YouTube/@SerranoAcademy