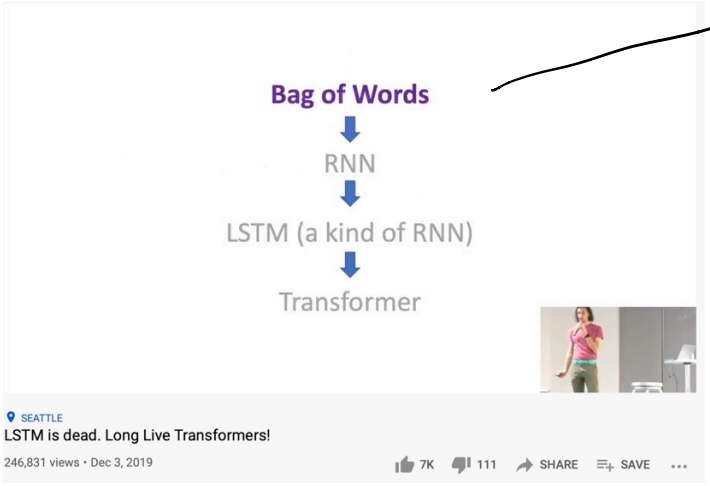
Challenges in working with natural language and sequential data

Challenges – words → vectors / numerical values / differentiable
pay attn to position
variable length input

Text generation / story generation
Translation
Creating Summaries
Question Answering
Text autocompletion
Irony/sarcasm detection
Grammar/spelling correction
Spam detection

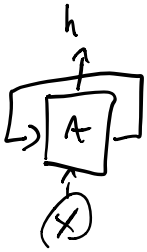
Google: GPT-3, DALL-E



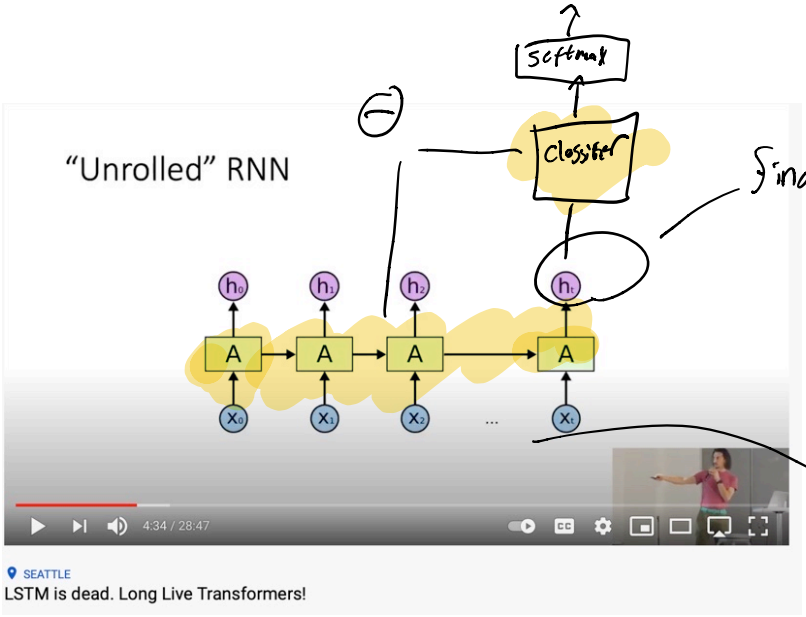
Input:
Vector of counts of each word in dict
the
cat
happy
...
$\in \mathbb{R}^{10000}$

"Work to live" vs "live to work"

RNN



Distrib over classes



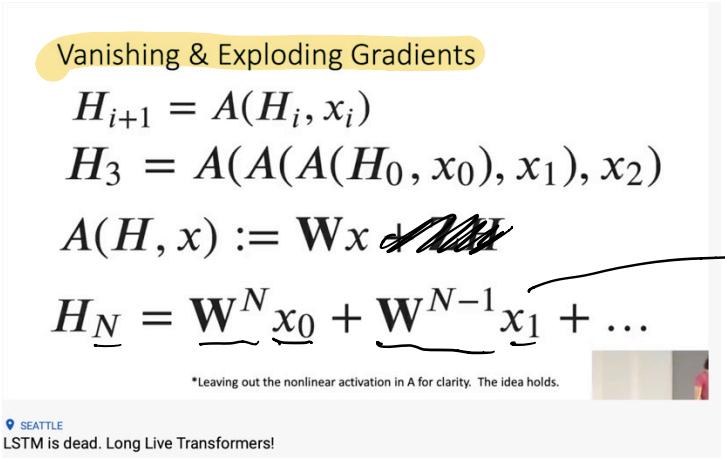
Data - {reviews, labels} y

θ

final state $\in \mathbb{R}^d$
put it into a classifier (FCN)

$f_\theta(x) = \hat{y}$

$\min_\theta \sum_i \ell(\hat{y}, y)$

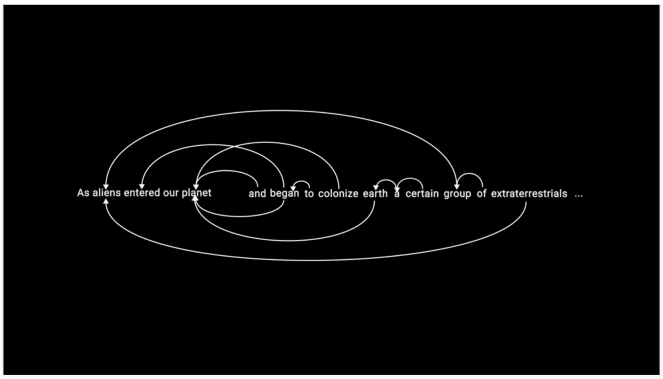


W^N when $N \to \infty$

if $W \in \mathbb{R}$,
$w > 1$, $w^N \to \infty$ "fast
$w < 1$, $w^N \to 0$

RNNs have a limited window of past times/positions that They can pay attention to

Attention

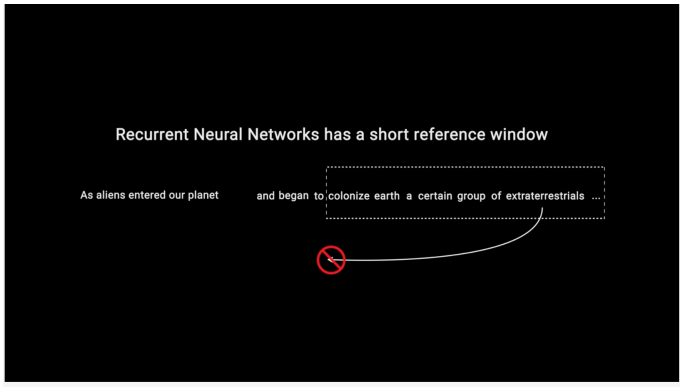


Illustrated Guide to Transformers Neural Network: A step by step explanation

63,343 views • Apr 27, 2020 2.9K 18 SHARE SAVE

The A.I. Hacker - Michael Phi
16.2K subscribers

SUBSCRIBE



Illustrated Guide to Transformers Neural Network: A step by step explanation

63,343 views • Apr 27, 2020 2.9K 18 SHARE SAVE

The A.I. Hacker - Michael Phi
16.2K subscribers

SUBSCRIBE

Visualizing Attention

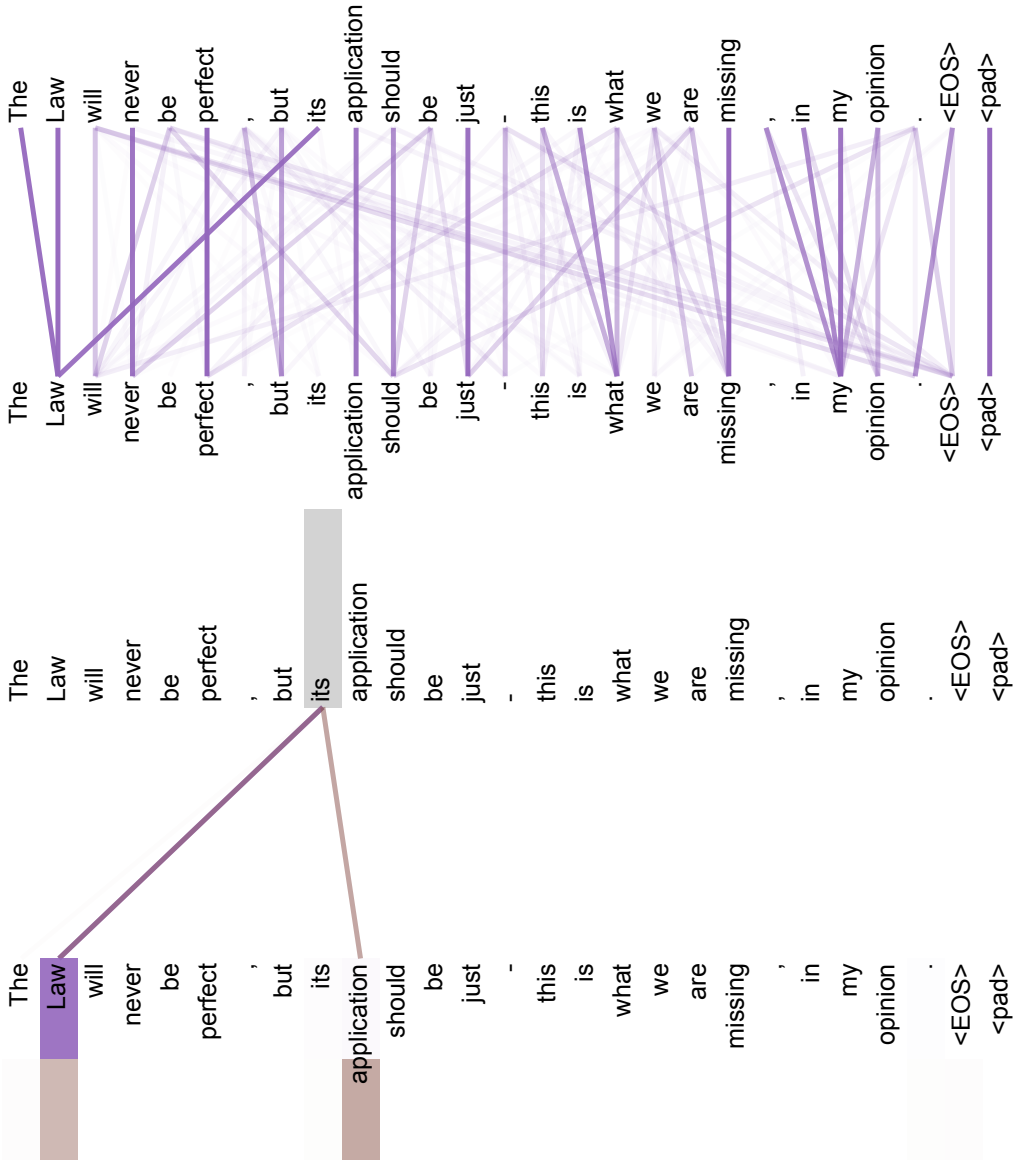
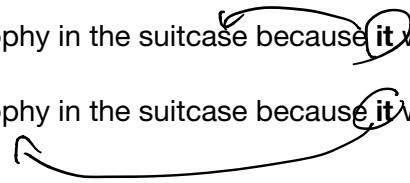


Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top: Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5 and 6. Note that the attentions are very sharp for this word.

Example of challenges that attention can help with: Machine Translation

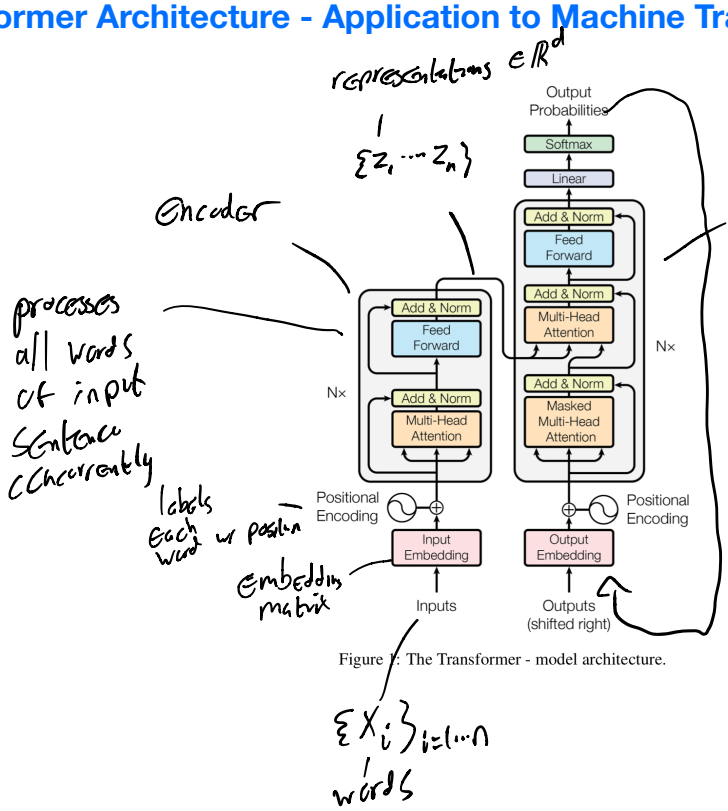
I didn't put the trophy in the suitcase because **it** was too **small**

I didn't put the trophy in the suitcase because **it** was too **big**



Transformer Architecture - Application to Machine Translation

D



Encoder: entire sentence is passed in (n tokens), it returns a representation for each input token. That representation has built in the notion of attention.

Decoder: generate one token at a time (in an autoregressive manner) in the target language

Figure 1: The Transformer - model architecture.

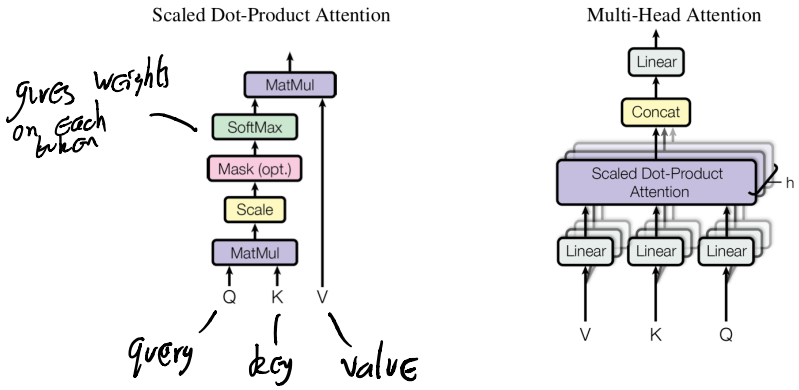
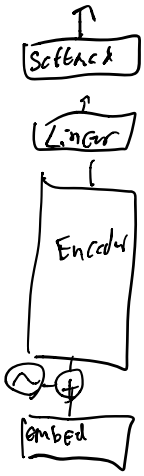


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Application to other NLP tasks

If you wanted to do sentiment analysis how would you modify this architecture?



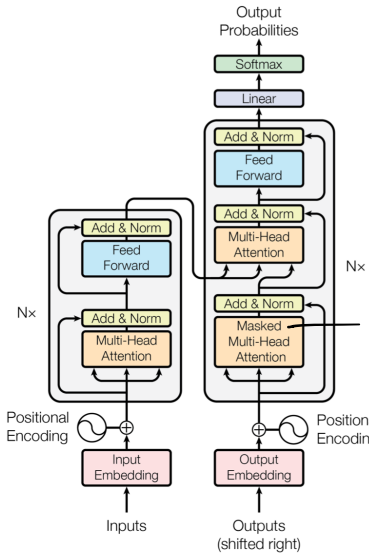
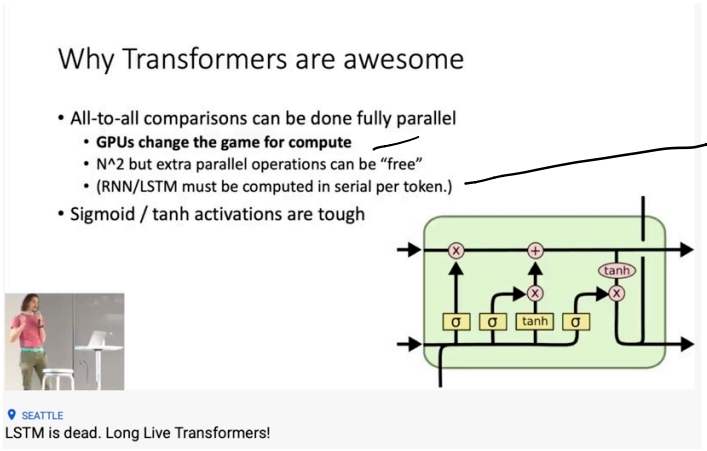


Figure 1: The Transformer - model architecture.

What architecture would I use for unconditional text generation?

use the decoder

can only attend to words gen. already



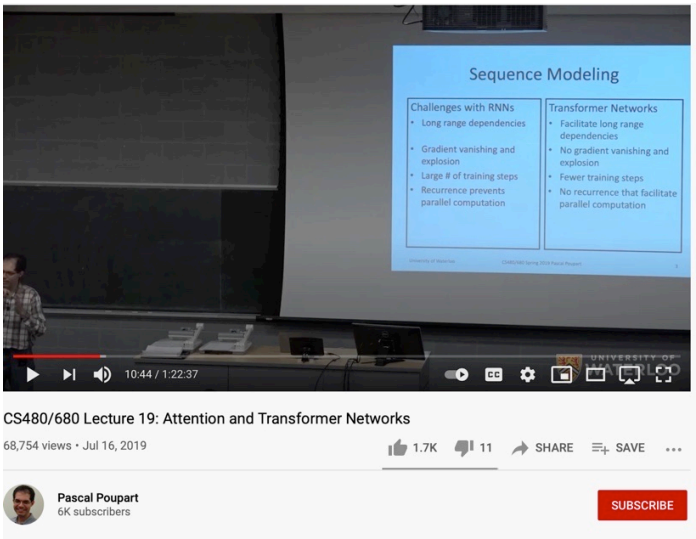


Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$