Intro to NLP

> *... reviewing machine learning literature about the common Natural Language Processing (NLP) models*

- https://www.kdnuggets.com/2020/01/guide-natural-language-generation.html

Attention is All You Need

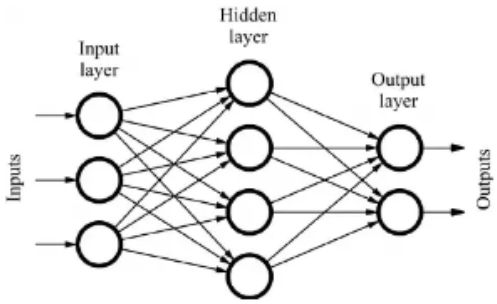
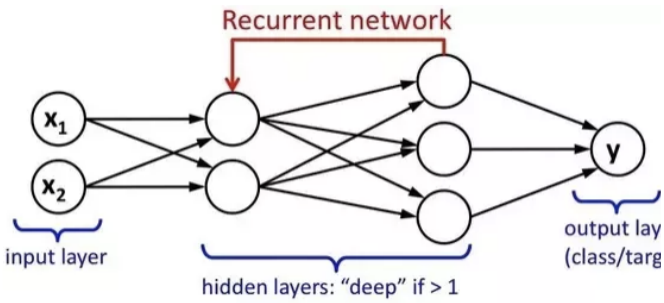
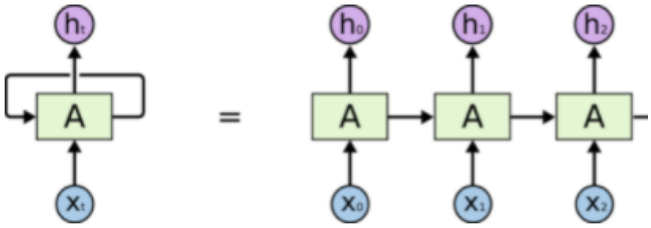
- Goal : propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely
- Link : https://research.google/pubs/pub46201/ (12/2017)
- Code : https://github.com/tensorflow/models/tree/master/official/nlp/nhnet
- Credible source: Google Research

Model

development of NLP mechanism	description

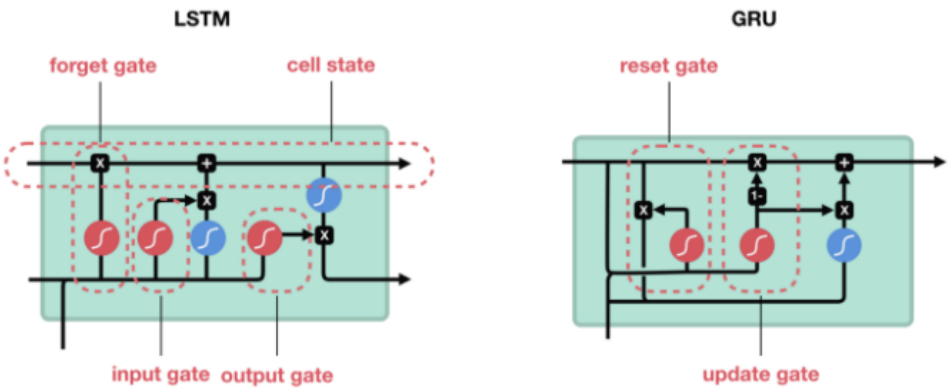
Recurrent Neural Network (RNN) - 1986

- a class of neural networks that is not feed-forward
- long-term dependency problem
  - vanishing gradients
  - hampers learning of long data sequences

feed-forward	recurrent
	  An unrolled recurrent neural networ
• feed-forward neural network • only feed the input in the forward direction • output from one unit is only fed to units further ahead	• there are loops in the network graph, • and the output of one unit may go back to one of the already visited u • allowing information to be passed from one step of the network to the

Long Short Term Memory (LSTM) - 1997

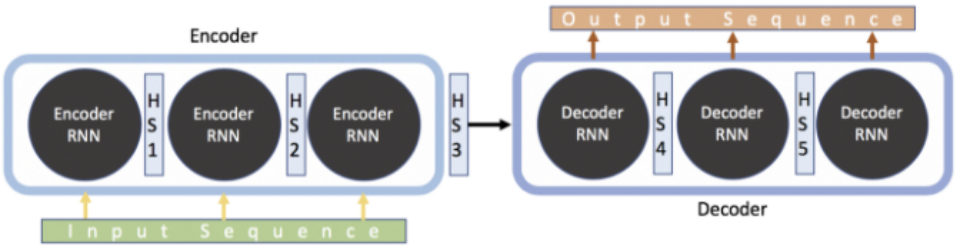
- a special type of RNNs
- include a 'memory cell' that can maintain information in memory for long periods of time
- remove or add information to the cell state by gates
- each gate can have value between zero and one
- https://guillaumegenthial.github.io/sequence-to-sequence.html
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/



- LSTM - cell state, and three gates to control cell state
- Gated Recurrent Unit(GRU) - two gates for reset and update

Seq2Seq - 2014

- a problem setting, where input is a sequence and output is also sequence
- example of sequence-to-sequence problems are machine translation, question answering, generating natural language description of videos, automatic summarisation, etc.

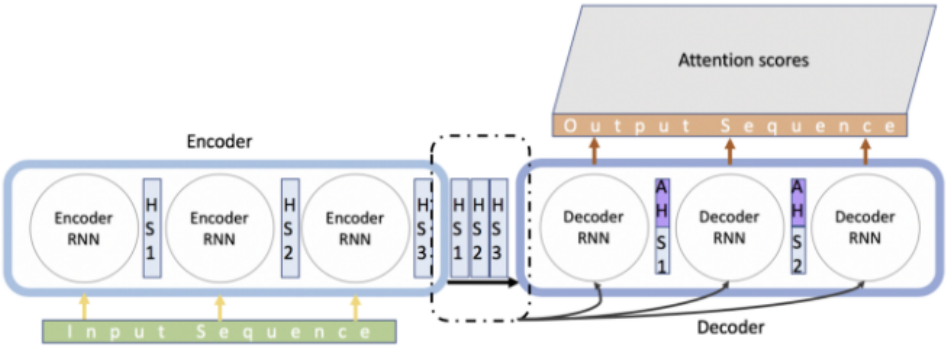


- seq2seq models that use LSTMs or RNNs as modules inside them, where a sequence-to-sequence model is just a model that works for sequence to sequence tasks
- using encoder-decoder paradigm - encoder captures the context of the input sequence in the form of a hidden state vector and send it to the decoder which then produces the output sequence
- bottleneck problem - due to the sequential order of word processing, it's harder for the context vector to capture all the information contained in a sentence for long sentences with complicated dependencies between words
- https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263

- https://guillaumegenthial.github.io/sequence-to-sequence.html

Seq2Seq with attention mechanism

- previously, output of the encoder was one vector, now have a matrix composed by each of the hidden states
- decoder know which part of the matrix to focus on with attention scores
- attention score - the alignment model scores how well an input (represented by its hidden state) matches with the previous output (represented by attention hidden state) and does this matching for every input with the previous output. Then a softmax is taken over all these scores and the resulting number is the attention score for each input



Seq2Seq Attention Based Model

- attention hidden vector - context vector which is weighted sum of the input hidden states combined with the hidden state vector

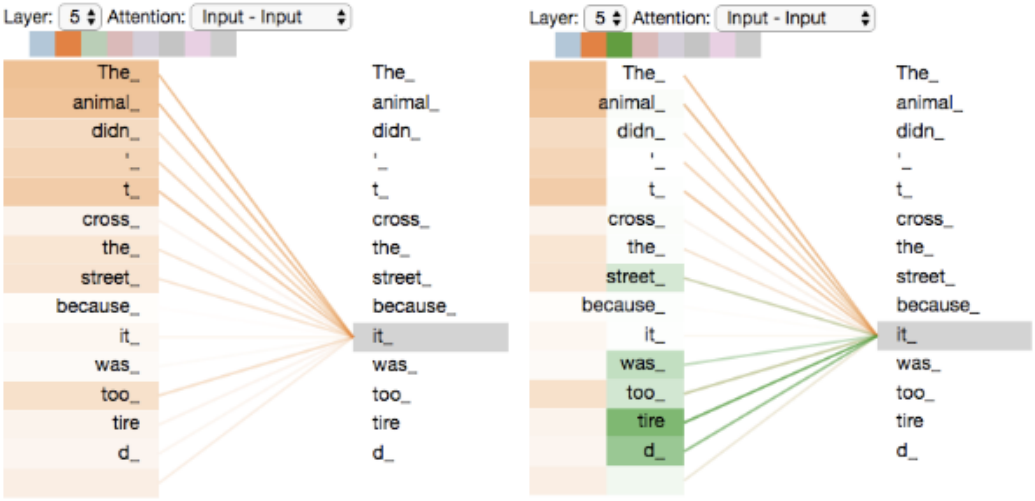
- context vector - combined with the hidden state vector by concatenation and this new attention hidden vector is used for predicting the output at that time instance. Note that this attention vector is generated for every time instance in the output sequence and now replaces the hidden state vector

- a solution to the bottleneck problem - allows the model to focus on different parts of the input sequence at every stage of the output sequence allowing the context to be preserved from beginning to end
- https://towardsdatascience.com/what-is-attention-mechanism-can-i-have-your-attention-please-3333637f2eac

Attention - 2015

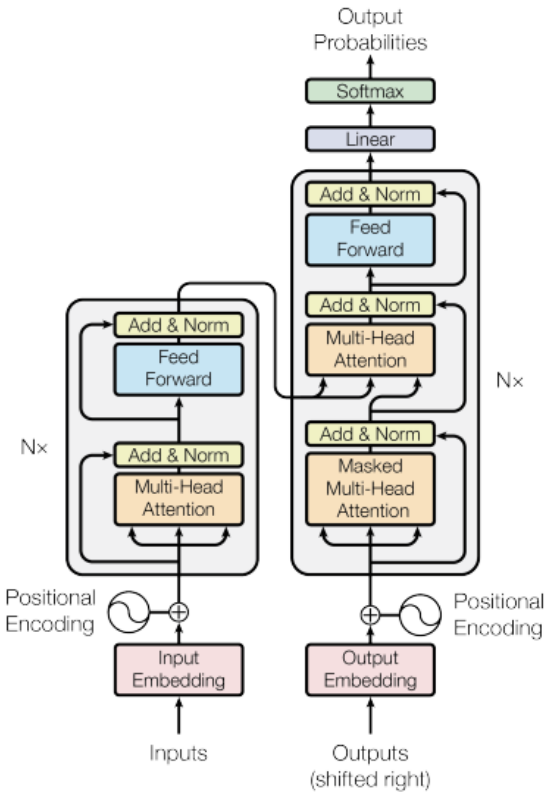
- give larger weights for the more informative parts
- self-attention
  - how each word affects to other words in one sentence
  - compute dependency relationships between words of the same sentence

- self-attention and multi-head attention



▪ multi-head attention ▪ the concept of adding dimensions or subspaces to the self-attention mechanism to retrieve ▪ https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html ▪ https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/ ▪ https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/	
Transformer - 2017 • **the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution**	

- encoder
  - composed of a stack of identical layers
  - each layer has two sub-layers
  - the first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network
  - employ a residual connection around each of the two sub-layers, followed by layer normalisation



- decoder
  - also composed of a stack of identical layers
  - encoder-decoder attention - performs multi-head attention over the output of the last layer of encoder stack contextual relationship of each word in source text
  - masking - to make sure that previous token are attended as prediction of next token only depends on its previous tokens
- https://www.linkedin.com/pulse/rise-transformers-imtiaz-adam
- https://jalammar.github.io/illustrated-transformer/

Bidirectional Encoder Representations from Transformers (BERT) - 2018 • a technique for natural language processing pre-training developed by Google • model with encoders (bidirectional self-attention heads) • use of transformer , an attention mechanism that learns contextual relations between words in a text • as opposed to directional models, which read the text input sequentially, the transformer encoder reads the entire sequence of words at once • allows the model to learn the context of a word based on all of its surroundings • https://github.com/google-research/bert • https://nlp.stanford.edu/seminar/details/jdevlin.pdf	

Generative Pre-trained Transformer 3 (GPT-3) - 2020 • autoregressive language model that uses deep learning to produce human-like text • predicts the probability of a given sentence existing in the world • trained on an unlabelled dataset using the Common Crawl and Wikipedia with a random removal of words leaving the model to learn to fill the gaps by application of solely the neighbouring words used as contex • https://osf.io/m6gcn	

Experiment

training data	experiment	method	evaluation	output
standard WMT 2014 1. English-German dataset consisting of about 4.5 million sentence pairs 2. English-French dataset consisting of 36M sentences and split tokens into a 32000 word-piece vocabulary ▪ WMT 2014 is a collection of datasets used in shared tasks of the Ninth Workshop on Statistical Machine Translation ▪ http://www.statmt.org/wmt14/index.html	▪ training models ▪ base model - trained 100,000 steps for 12 hours ▪ big model - trained 300,000 steps for 3.5 days ▪ one machine with 8 NVIDIA P100 GPUs ▪ optimiser - adam optimiser ▪ regularisation - residual dropout(0.1) and label smoothing (0.1) ▪ and compare BLEU scores and training cost with previous models such as ByteNet, Deep-Att + PosUnk, GNMT + RL, ConvS2S, and MoE	transformer	• BLEU (Bilingual Evaluation Understudy) • algorithm for evaluating the quality of text which has been machine-translated from one natural language to another • the closer a machine translation is to a professional human translation, the better it is • BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics • scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations • those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality • https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german • corpus is a language resource consisting of a large and structured set of text	▪ English-German translation ▪ big model outperforms by more than 2.0 BLEU, having 28.4 scores ▪ base model even surpasses all previously published models ▪ English-French translation ▪ big model achieves 41.0 BLEU score and less than 1/4 training cost

ELMo (Embeddings from Language Models)

- contextualised word-embeddings
  - instead of using a fixed embedding for each word, ELMo looks at the entire sentence before assigning each word in it an embedding
  - embedding word based on the context it's used in to both capture the word meaning in that context as well as other contextual information
- uses a bi-directional LSTM trained on a specific task to be able to create those embeddings
- gains its language understanding from being trained to predict the next word in a sequence of words
- trained on a massive dataset that such a model can learn from without needing labels