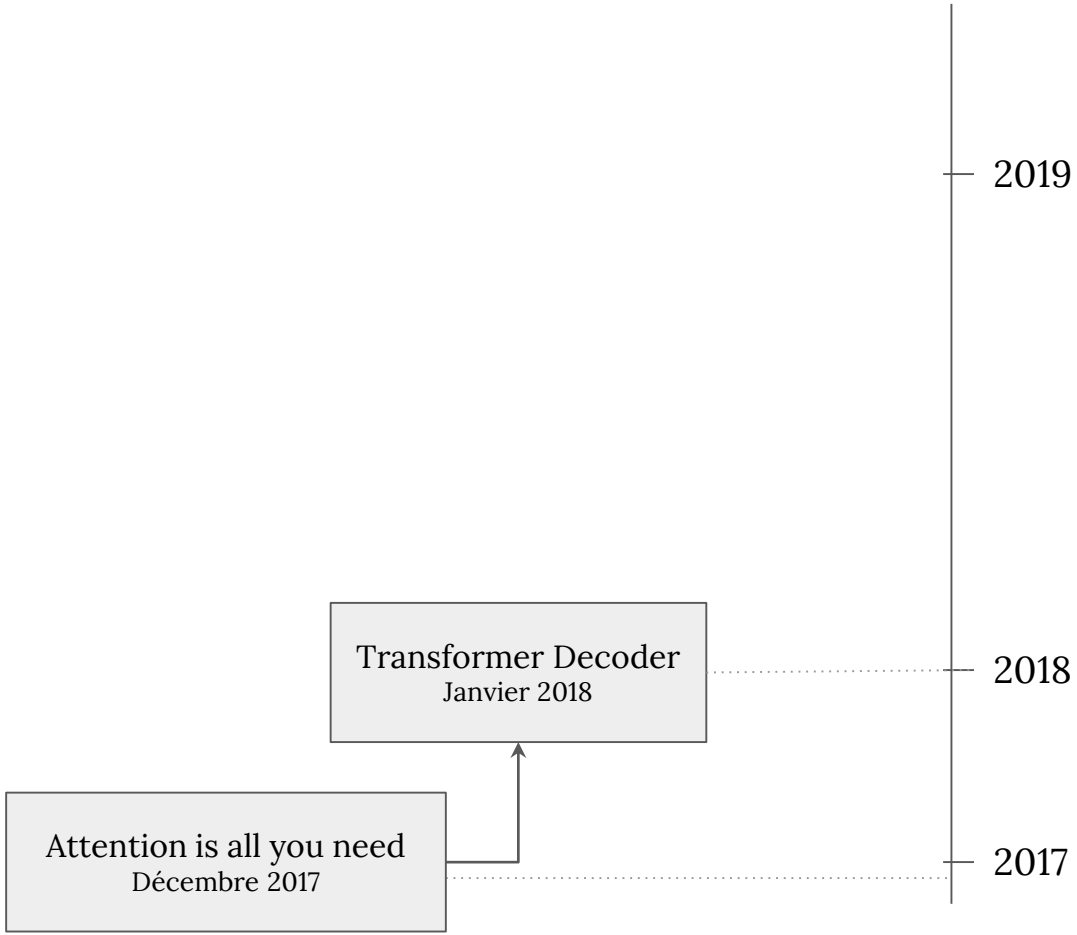
Transformer-Decoder

Generating Wikipedia by Summarizing Long Sequences
Peter J. Liu et al.


2019
Transformer Decoder
Janvier 2018
2018
Attention is all you need
Décembre 2017
2017

Outline

- They consider the task of multi-document summarization where multiple documents are distilled into a single summary.
- Introduces the decoder-only architecture thats scales to longer sequences than the encoder-decoder architecture.

Dataset

Deep learning

From Wikipedia, the free encyclopedia

For deep versus shallow learning in educational psychology, see Student approaches to learning. For more information, see Artificial neural networ

Deep learning (also known as **deep structured learning** or **hierarchical learning**) is part of a broader family of machine learning methods based on representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.[1][2][3]

etworks and recurrent neural networks have been applied to fields including c
ork filtering, machine translation, bioinformatics, drug design, medical image a
s comparable to and in some cases superior to human experts.[4][5][6]

d communication patterns in biological nervous systems yet have various diffe
n brains), which make them incompatible with neuroscience evidences.[7][8][9]

References [edit]

1. ^ a b c d e f Bengio, Y.; Courville, A.; Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **35** (8): 1798–1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50. PMID 23787338.
2. ^ a b c d e f g h Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". *Neural Networks*. **61**: 85–117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003. PMID 25462637.
3. ^ Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). Learning". *Nature*. **521** (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature1453 PMID 26017442.
4. ^ a b Ciresan, Dan; Meier, U.; Schmidhuber, J. (June 2012 column deep neural networks for image classification".
66. ^ Deng, L.; Abdel-Hamid, O.; Yu, D. (2013). "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion" (PDF). *Google.com*. ICASSP.
67. ^ a b Sainath, T. N.; Mohamed, A. r; Kingsbury, B.; Ramabhadran, B. (May 2013). "Deep convolutional neural networks for LVCSR". *2013 IEEE International Conference on Acoustics, Speech and Signal*

Deep Learning | Coursera
https://fr.coursera.org › Parcourir › Science des données › Apprentissage automatique
Learn **Deep Learning** from **deeplearning**.ai. If you want to break into AI, this Specialization will help you do so. **Deep Learning** is one of the most highly sought ...

tion and
ganizers: Li

À la une


Le prix Turing à trois pionniers du deep learning
ICTjournal · Il y a 2 jours


Machine Learning : le futur de l'IA est déjà là
Contrepoints · Il y a 11 heures

Plus de résultats pour "Deep Learning"

etworks.

Deep Learning
deeplearning.net/ ▾ Traduire cette page
Deep Learning is a new area of **Machine Learning** research, which has been introduced with the objective of moving **Machine Learning** closer to one of its ...

Model

- Extractive stage
  - Relevant sentence extraction
- Abstractive stage
  - Wikipedia article generation

Extraction stage

- Given ***C*** the cited sources and ***S*** the search results
- For each article in (***C***, ***S***), create a ranked list of of paragraphs
  - There is a couple of methods to do this (identify a trivial baseline, tf-idf, etc.)
- Concatenate all the ranked paragraphs and extract the ***L*** tokens, ***L*** being typically of length 11000

Abstractive stage

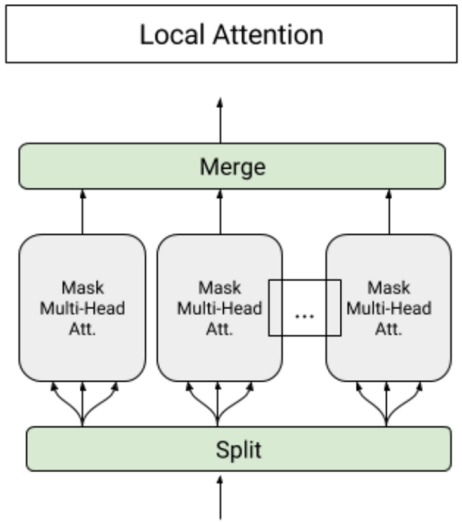
- Given the sequence of words of length ***L***
  - Modify the Transformer-Encoder-Decoder (T-ED) as Transformer-Decoder (T-D) only, **similar architecture (5 instead of 6 layers)**
  - $(m^1, ..., m^n)$ -> $(y^1, ..., y^n)$ (transducer model) becomes
  - $(m^1, ..., m^n, \delta, y^1, ..., y^n)$ where δ is a special separator token
- They train the model as a traditional language model
- They **suspect** (would have been interesting to see concrete results!) that for monolingual text-to-text tasks redundant information is re-learned about language in the encoder and the decoder

Abstractive stage

- Introduction of:
  - Local Attention
  - Memory compressed attention

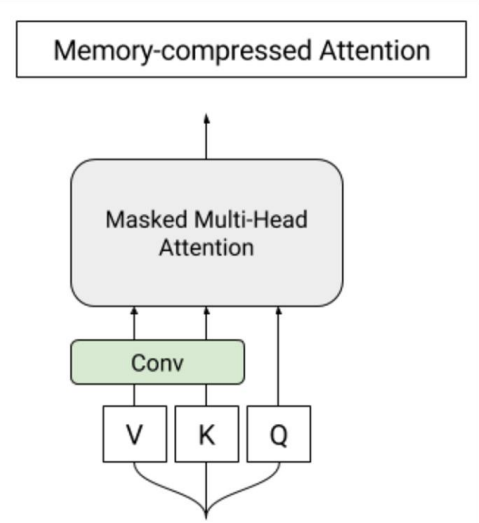
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Abstractive stage



Splits the sequence into individual smaller sub-sequences. The sub-sequences are then merged together to get the final output sequence.

Abstractive stage

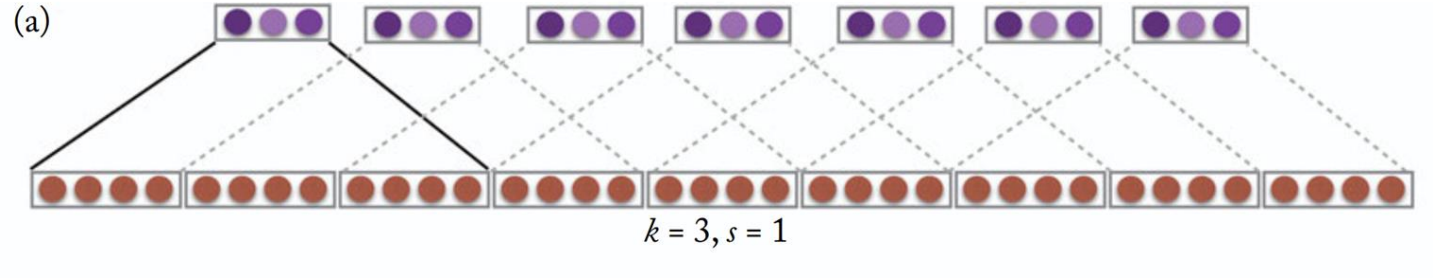


Reduces the number of keys and values by using a strided convolution (k=3, s=3).
The number of queries remains unchanged.

In contrast to local attention layers, which only capture the local information within a block, the memory compressed attention layers are able to exchange information globally on the entire sequence.

Compression with strided convolution

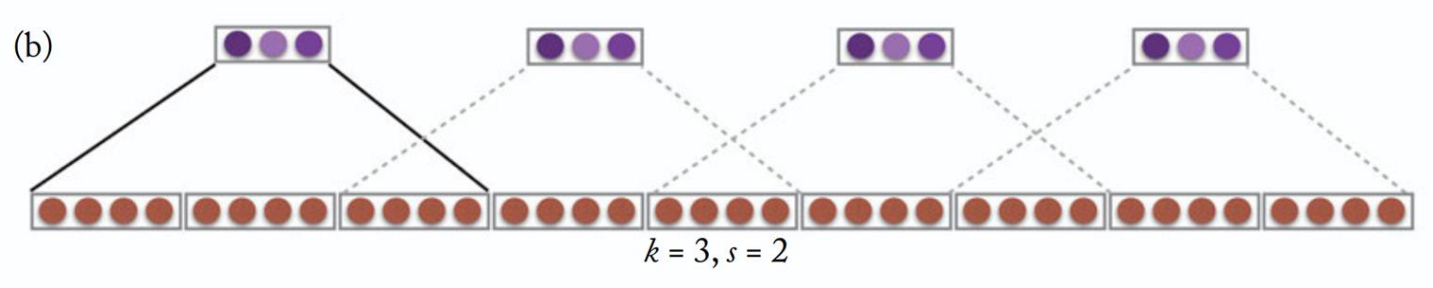
Output = 7



Input = 9

Compression with strided convolution

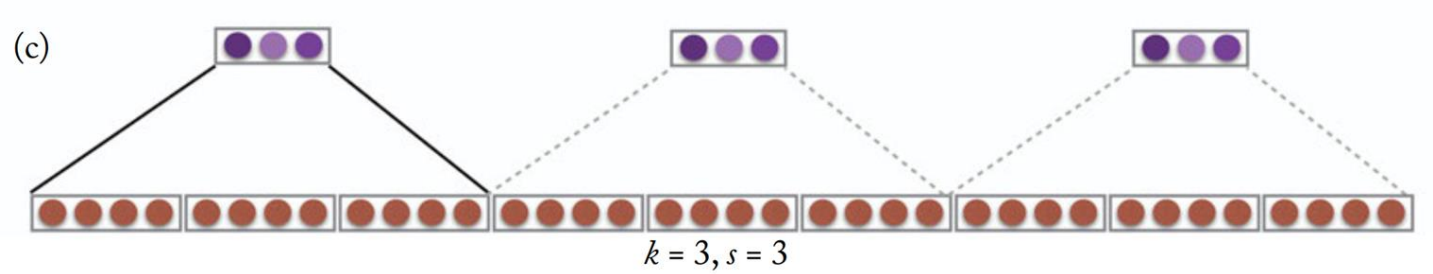
Output = 4



Input = 9

Compression with strided convolution

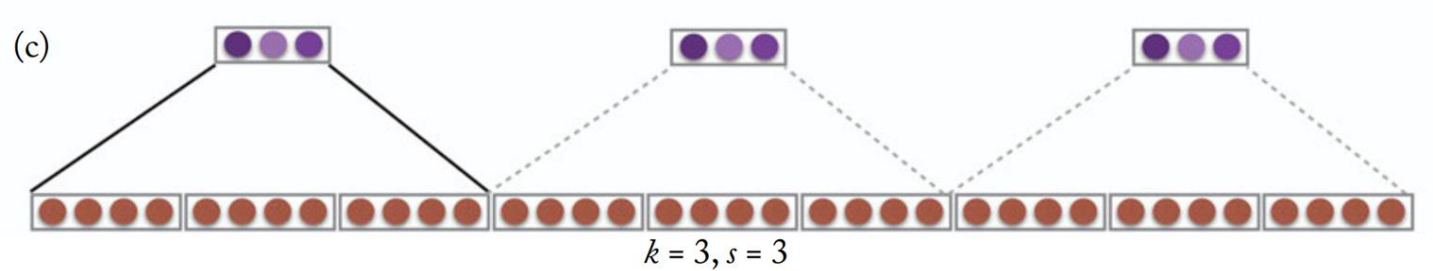
Output = 3



Input = 9

Compression with strided convolution

Output = 3



Input = 9

Allows to process sequences 3X longer!

Final architecture

- Combines local-attention and memory compressed attention on 5 layers:
  - Local-Compressed-Local-Compressed-Local

Results

- T-ED is able to learn from sequences around 500-1000 tokens
- T-D is able to learn from sequences of 4000 tokens before running out of memory
    - Adding Memory Compressed attention, improved performances with sequences of up to 11000 tokens

Results

Table 5: Linguistic quality human evaluation scores (scale 1-5, higher is better). A score significantly different (according to the Welch Two Sample t-test, with $p = 0.001$) than the *T-DMCA* model is denoted by *.

Model	**Focus**	**Grammar**	**Non-redundancy**	**Referential clarity**	**Structure and Coherence**
T-DMCA (best)	4.5	4.6	4.2	4.5	4.2
tf-idf-only	3.0*	3.6*	3.9	3.2*	2.7*
seq2seq-attention	3.0*	3.4*	2.1*	3.4*	2.3*

Conclusion

Differences with Attention is All you Need

- Remove the encoder architecture
  - By introducing a special *separator token*
- Use a memory compressed attention mechanism which allows to handle longer sequences