Long Range Attention and Visualizing BERT

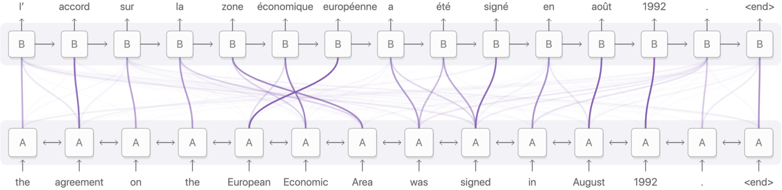
Presenter: Jack Lanchantin

University of Virginia
https://qdata.github.io/deep2Read/

201906

Outline

Attention



l' accord sur la zone économique européenne a été signé en août 1992 . <end>
B B B B B B B B B B B B B B B
A A A A A A A A A A A A A A A
the agreement on the European Economic Area was signed in August 1992 . <end>

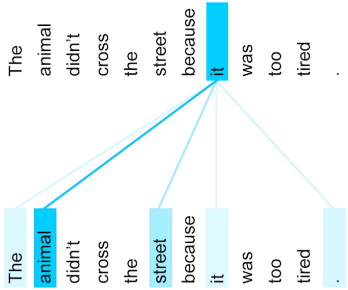
Attention

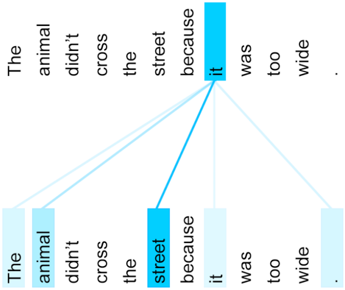
by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Self Attention





Self Attention
[7]

$$\alpha_{ij} = \frac{\exp\left(x_i^T x_j\right)}{\sum_{l=1}^{n} \exp\left(x_i^T x_l\right)} \tag{1}$$

$$x_i^{l+1} = \sum_{j=1}^{n} \alpha_{ij} x_j \tag{2}$$

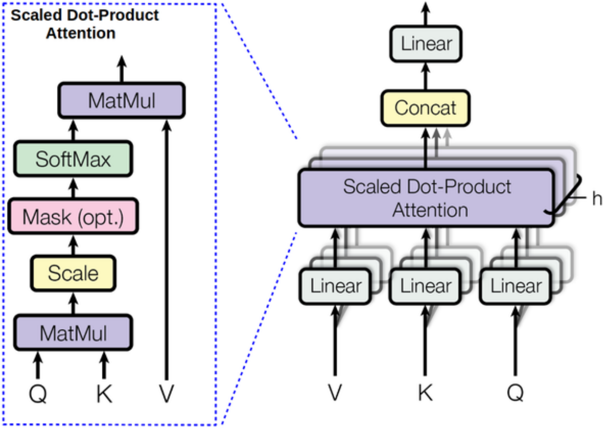
Self Attention
[7]

$$X^{l+1} = \mathrm{Attn}(X^l, X^l, X^l) \tag{3}$$

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

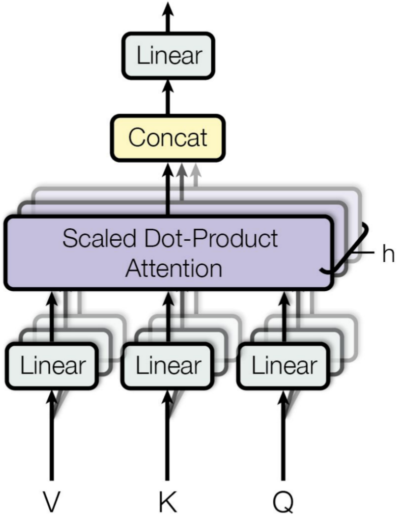
$$X^0 = \mathrm{lookupTable}(x) + \mathrm{positionEncoding}(x)$$

Self Attention
[7]

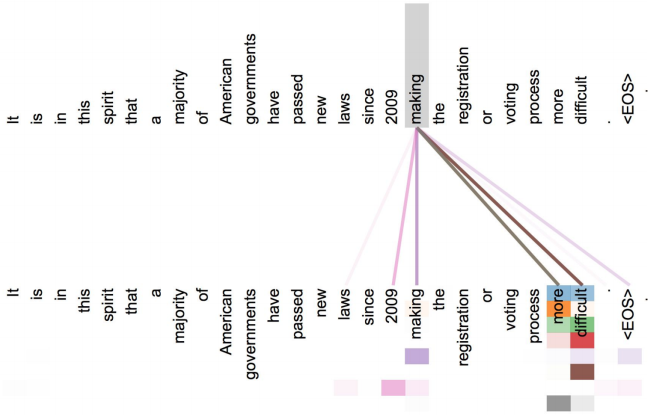


Multi-Head Attention

[7]

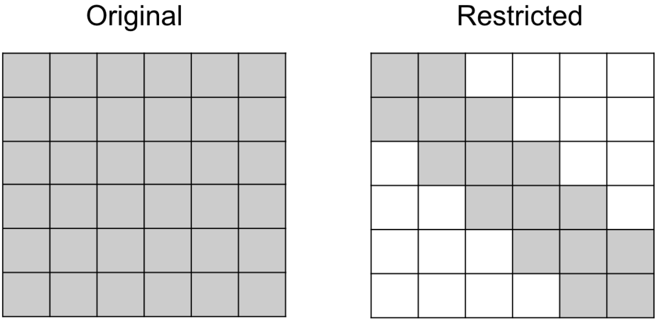


Multi-Head Attention
[7]



Restricted Neighbor Attention
[7]

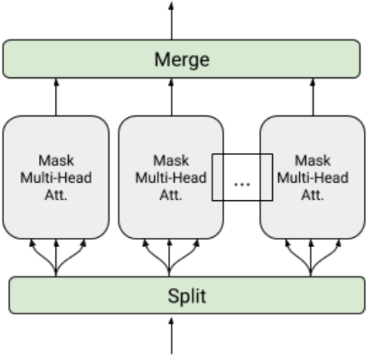
- Only allow attention to k neighbors



$$O(Nk)$$

Local Attention

[5]

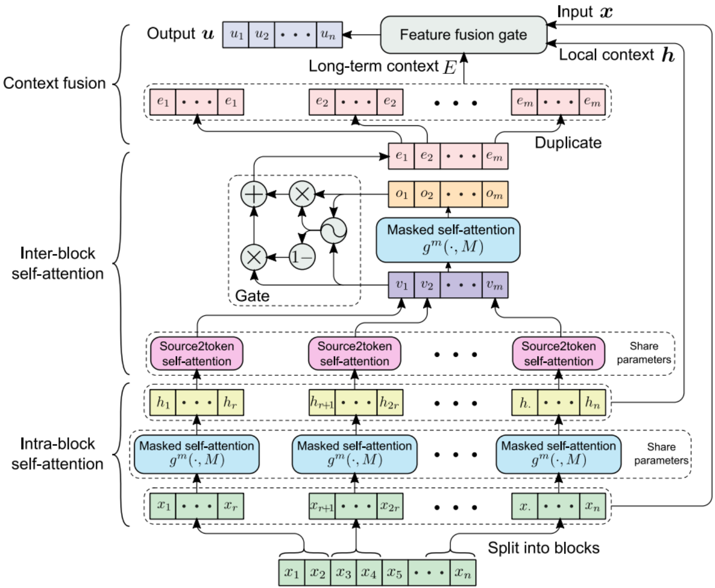


$$O(k^2)$$

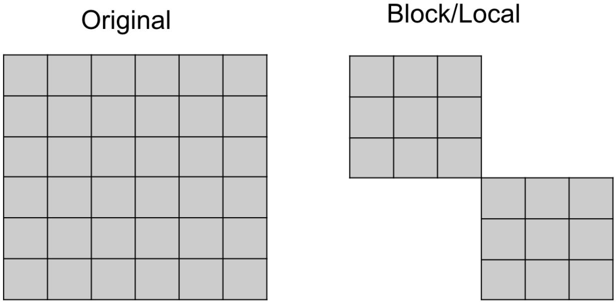
where k is the block size and $B = \frac{N}{k}$ is the number of blocks

Block Self Attention

[6]

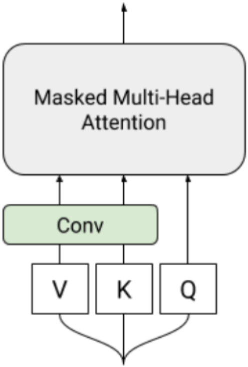


Local Attention



Memory Compressed Attention

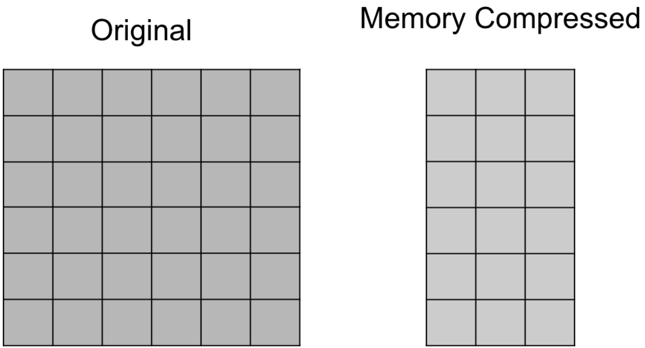
[5]



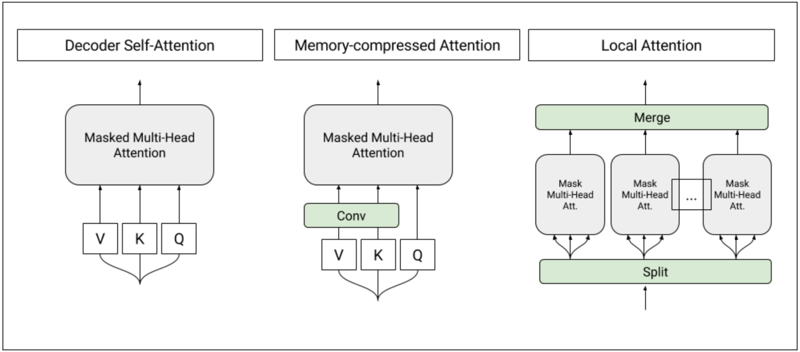
Reduce the number of keys and values by using a strided convolution. The number of queries remains unchanged.

$$O(N\frac{N}{k})$$

Memory Compressed



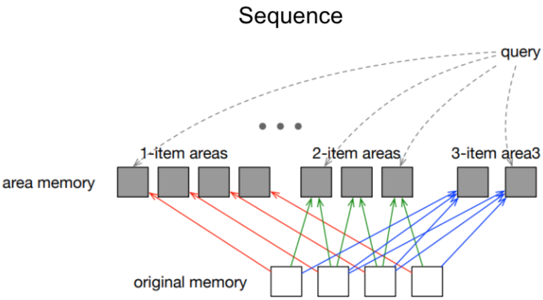
All Masks


Decoder Self-Attention
Memory-compressed Attention
Local Attention
Masked Multi-Head Attention
V
K
Q
Masked Multi-Head Attention
Conv
V
K
Q
Merge
Mask Multi-Head Att.
Mask Multi-Head Att.
...
Mask Multi-Head Att.
Split

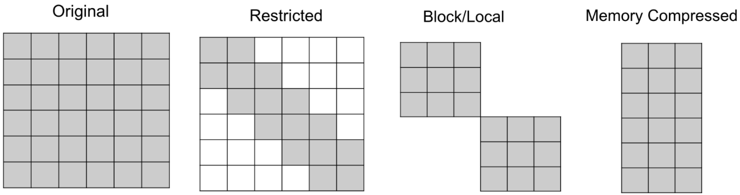
Area Attention

[4]

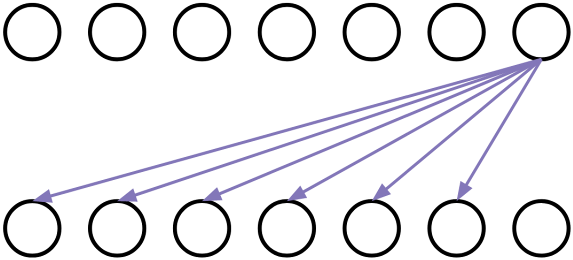
- Put groups of original memory keys (e.g. from individual tokens) into "areas"
  - Keys: mean of each area:
  - Values: sum of each area



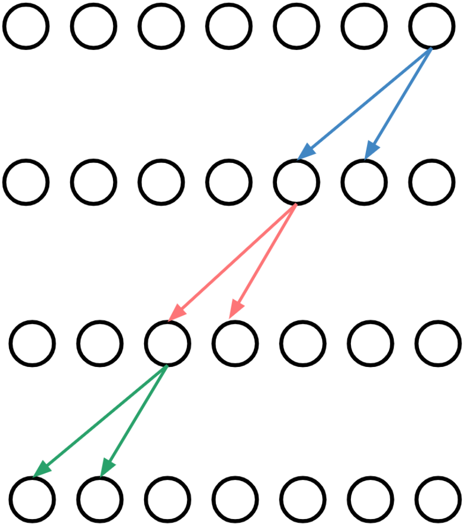
All Masks



Generating Long Sequences with Sparse Transformers [1]



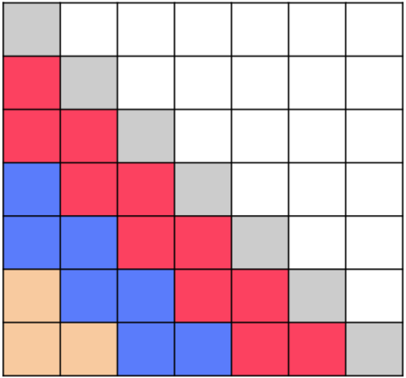
Generating Long Sequences with Sparse Transformers [1]



Generating Long Sequences with Sparse Transformers [1]

- Choosing p attention heads, set the attention width to $\sqrt[p]{N}$
- Reach full connectivity after p attention update steps
- Reduces effective computation to $O(N\sqrt[p]{N})$

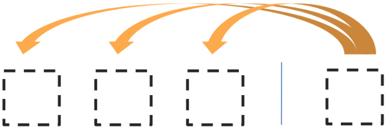
Generating Long Sequences with Sparse Transformers [1]



Generating Long Sequences with Sparse Transformers [1]

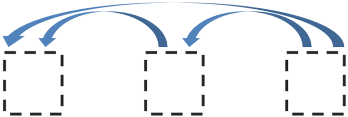
- S_i denotes the set of indices of the input vectors to which the embedding i attends
- Factorized self-attention instead has p separate attention heads, where the mth head defines a subset of the indices $A_i^{(m)} \subset \{j : j \leq i\}$ and lets $S_i = A_i^{(m)}$ where $|A_i^{(m)}| \propto \sqrt[p]{n}$
- For every $j \leq i$ pair, we set every A such that i can attend to j through a path of locations with maximum length $p + 1$. Specifically, if $(j, a, b, c, ..., i)$ is the path of indices, then $j \in A_a^{(1)}, a \in A_b^{(2)}, b \in A_c^{(3)}$ and so forth

Attention Types



Encoder-Decoder Attention

Encoder Self-Attention



MaskedDecoder Self-Attention

Music Transformer

[3]

$$\text{Relative Attention } = \text{Softmax}\left(\frac{QK^{\top} + S^{rel}}{\sqrt{D_h}}\right)V \qquad (5)$$

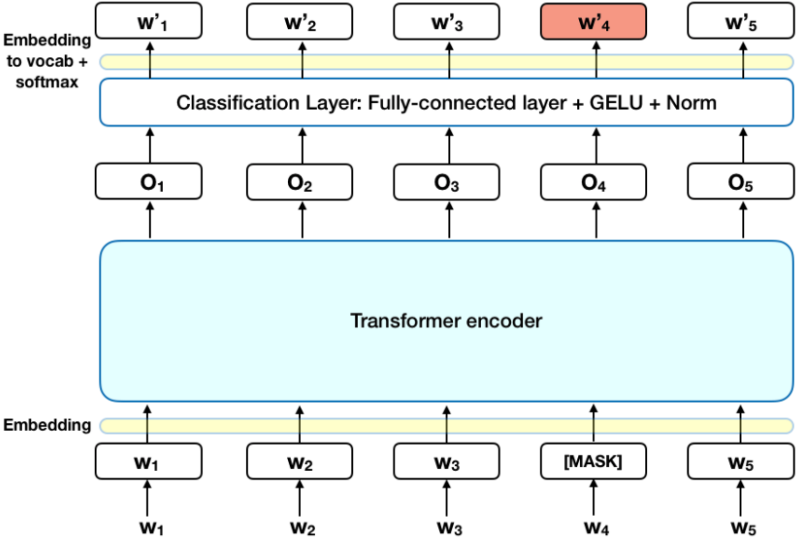
- S^{rel}, an $L \times L$ dimensional logits matrix which modulates the attention probabilities for each head.
- $S^{rel} = QR^{\top}$, where R is a tensor of shape (L, L, D_h) containing the embeddings that correspond to the relative distances between all keys and queries.

Outline

Transformers for Long Range Dependencies

Visualizing BERT

BERT



w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Context Embeddings

- Hewitt and Manning (2017) The authors find that after a single self attention step (before the nonlinearity) the square of the distance between context embeddings is roughly proportional to tree distance in the dependency parse.
- This paper seeks to answer why

Visualizing and Measuring the Geometry of BERT [2]

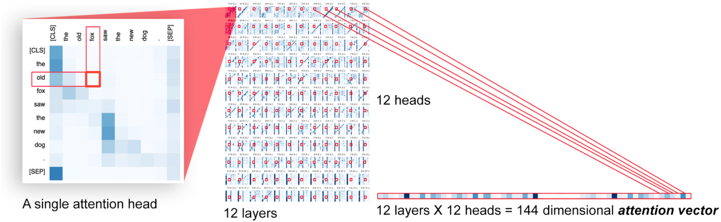
- Goal: explore BERT's internal representations
  - Investigate attention matrices
  - Investigate context embeddings in relation to parse trees
  - Find semantic representations of BERT embeddings

Semantics of Attention Matrices

- Attention matrices are built on relations between pairs of words. Do they represent grammar structure between these pairs?
- Formulation: can an attention vector for a pair of words classify a dependency relation?

Semantics of Attention Matrices

- Train linear model on the model-wide attention vector for pairs of words

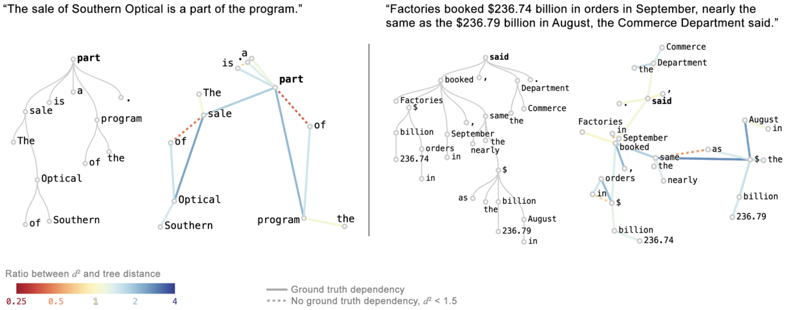


- 85.8% accuracy on dependency relation prediction from Penn Treebank
- i.e. syntactic information is encoded in attention vectors

Context Embeddings

- After a single self attention step, the square of the distance between context embeddings is roughly proportional to tree distance in the dependency parse tree
- Suggests that BERT embeddings are a good alternative to parse tree embeddings

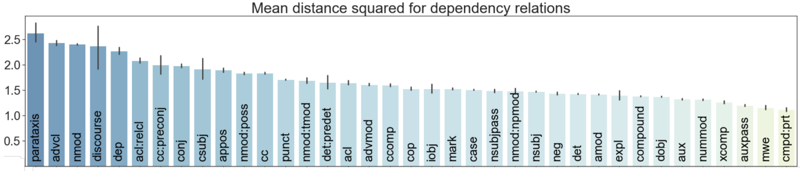
Context Embedding Relationships



Distance Between Words of All Relations

- Is the actual difference between embedding distance and the tree distance merely noise, or a more interesting pattern?
- By looking at the average embedding distances of each dependency relation, we can see they vary

Average Distance Between Words of all Relations



- Suggests that BERT's syntactic representation has an additional quantitative aspect beyond traditional dependency grammar

Contextual Semantics

- Does BERT actually encode contextual meaning into its representation
  - e.g. does "bark" refer to a tree or a dog

Contextual Semantics Visualization Tool

- **Input**: word
- **Retrieves**: 1000 sentences from wikipedia containing that word
- **Outputs**: clusters separating the embeddings of the input word for each sentence

Contextual Semantics Visualization Tool

Quantitative Semantic Evaluation

- For a given word with n senses, create a nearest-neighbor classifier where each neighbor is the centroid of a given word sense's BERT-base embeddings in the training data.
- To classify a new word we find the closest of these centroids
- State of the art F1 score of 71.1

Concatenated Similarity Ratio

- If word sense is affected by context, then we should be able to influence context embedding positions by systematically varying their context
- Idea: concatenate sentences of the same word with different semantic meanings

A: "He thereupon *went* to London and spent the winter talking to men of wealth." *went*: to move from one place to another.

B: "He *went* prone on his stomach, the better to pursue his examination." *went*: to enter into a specified state.

Concatenated Similarity Ratio

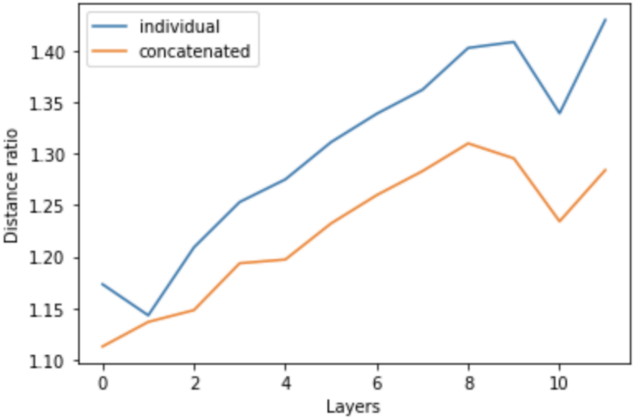


Figure 5: Average ratio of similarity to sense A vs. similarity to sense B.

References I

[1] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[2] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of bert. *arXiv preprint arXiv:1906.02715*, 2019.

[3] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, and Douglas Eck. An improved relative self-attention mechanism for transformer with application to music generation. *arXiv preprint arXiv:1809.04281*, 2018.

[4] Yang Li, Lukasz Kaiser, Samy Bengio, and Si Si. Area attention, 2019.

References II

[5] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[6] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Bi-directional block self-attention for fast and memory-efficient sequence modeling. *arXiv preprint arXiv:1804.00857*, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.