Triplet Attention: Rethinking the similarity in Transformers

Haoyi Zhou[1], Jianxin Li[1*], Jieqi Peng[1], Shuai Zhang[1], Shanghang Zhang[2]

[1]Beihang University [2]UC Berkeley

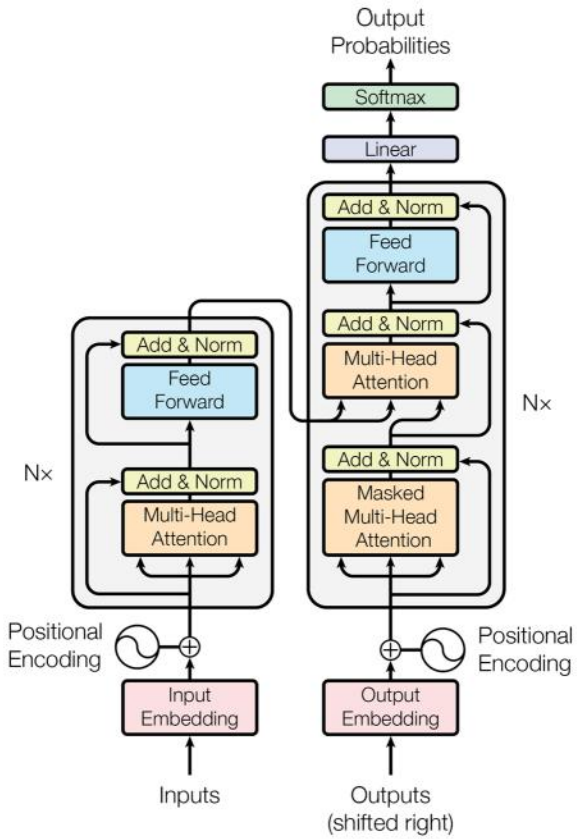
Content

- **Background**
- **The Triplet Attention**
- **Experimental Results**
- **Summary & Future Work**

Content

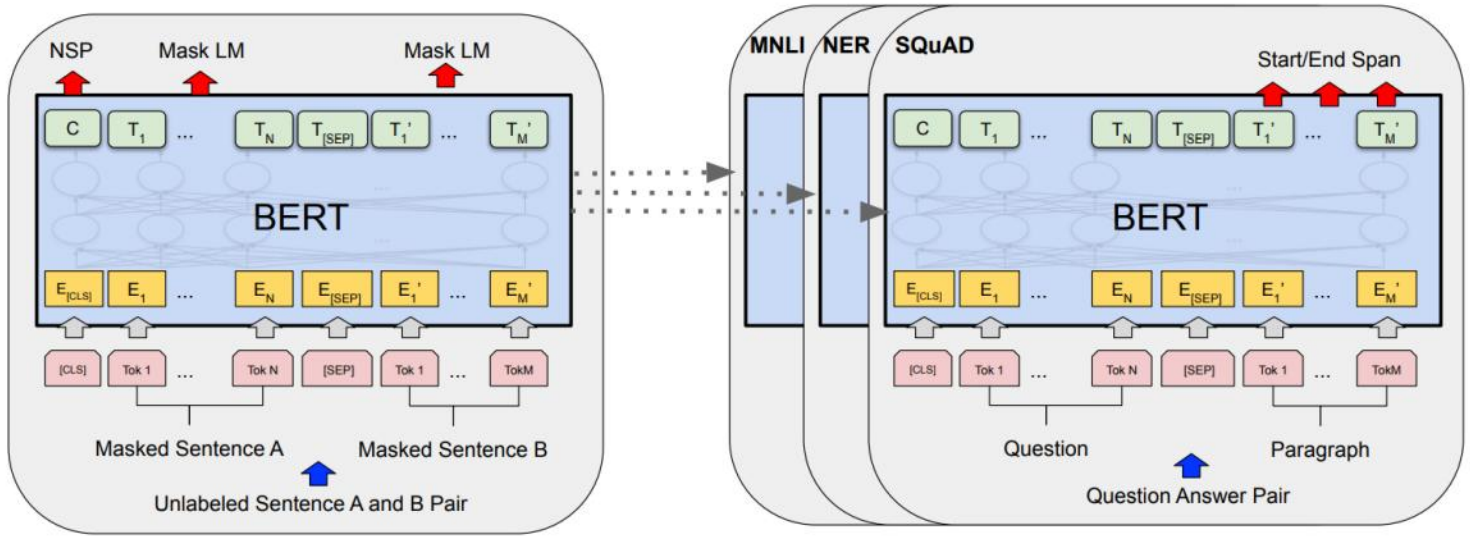
- **Background**
- **The Triplet Attention**
- **Experimental Results**
- **Summary & Future Work**

Transformer models' success



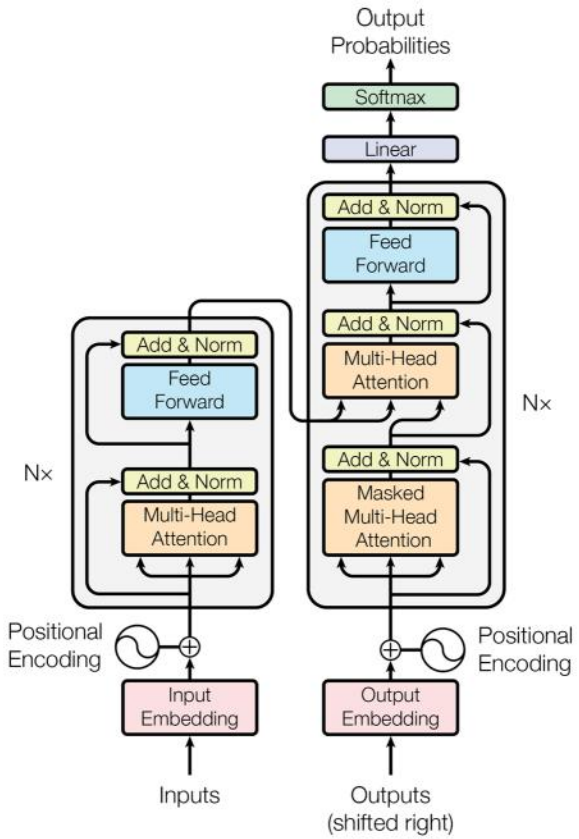
Transformer

Natural Language Processing



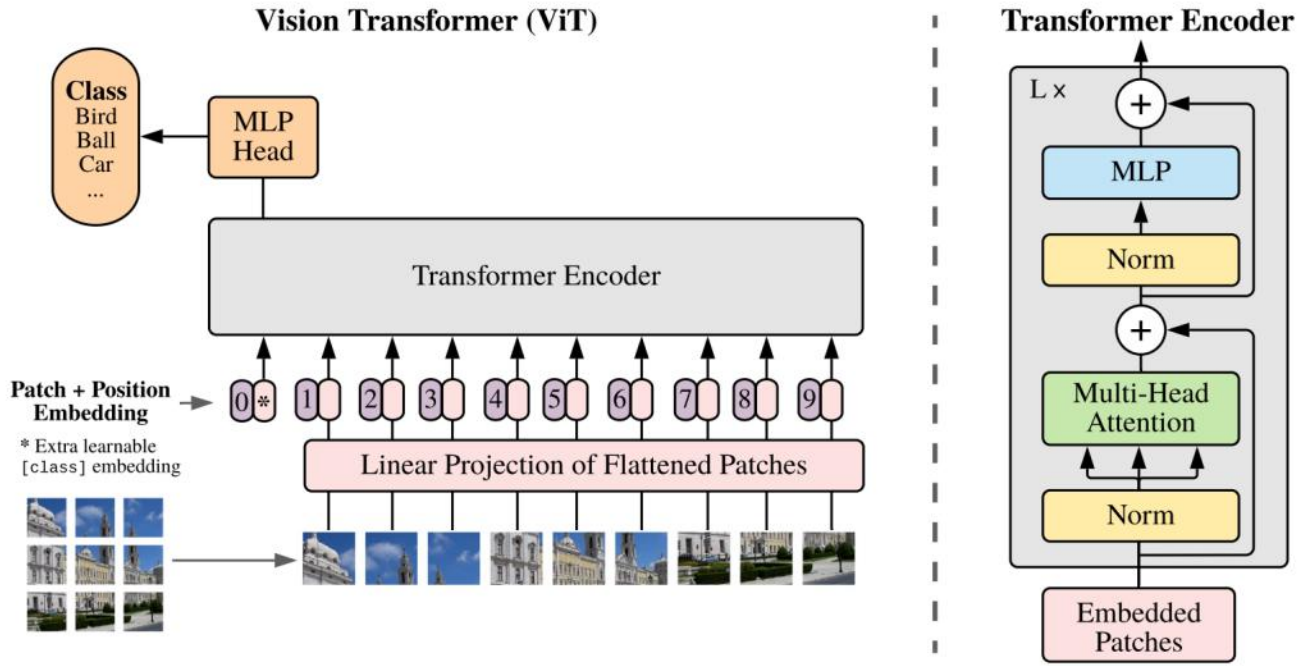
Bert

Transformer models' success



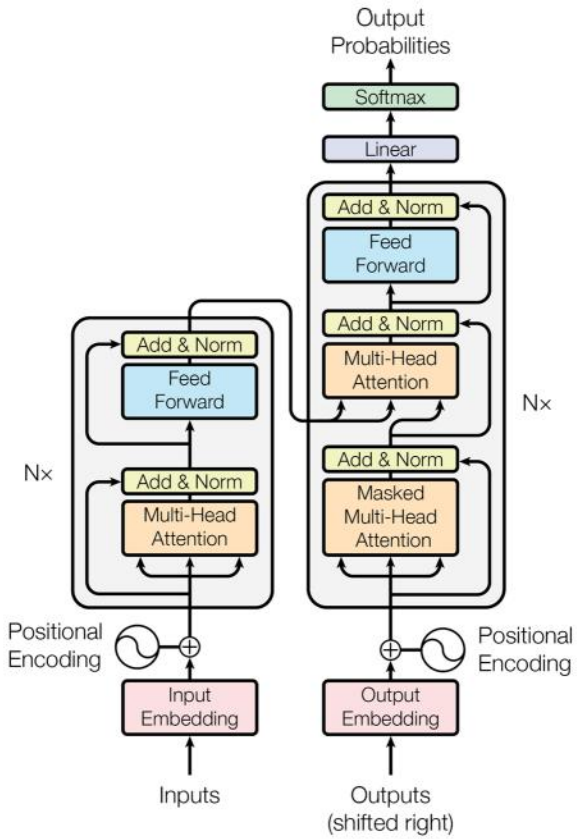
Transformer

Natural Language Processing → Computer Vision



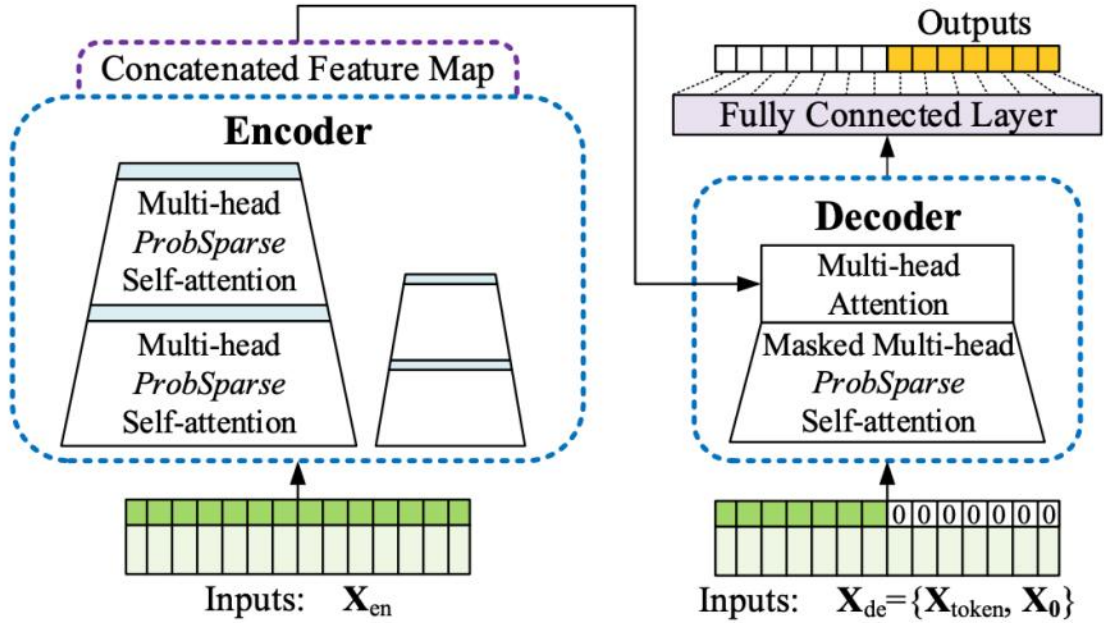
ViT

Transformer models' success



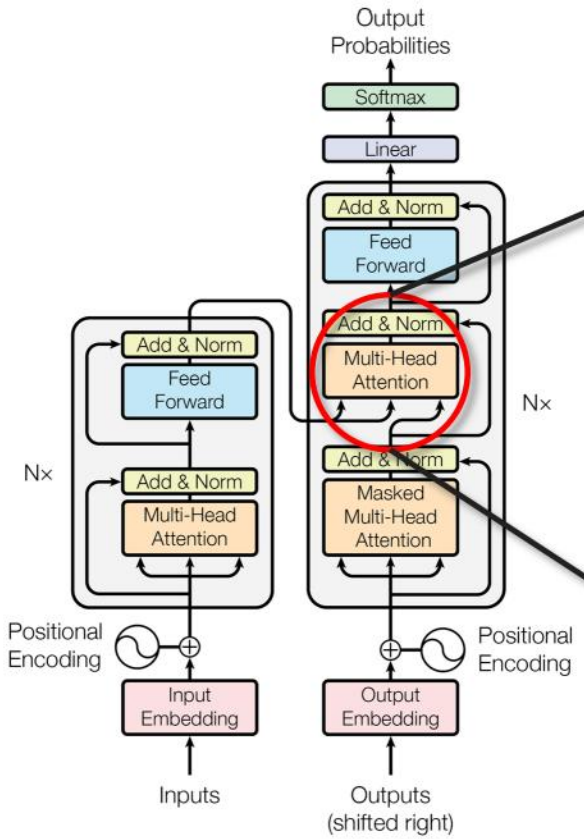
Transformer

Natural Language Processing | Computer Vision | Time-series Analysis



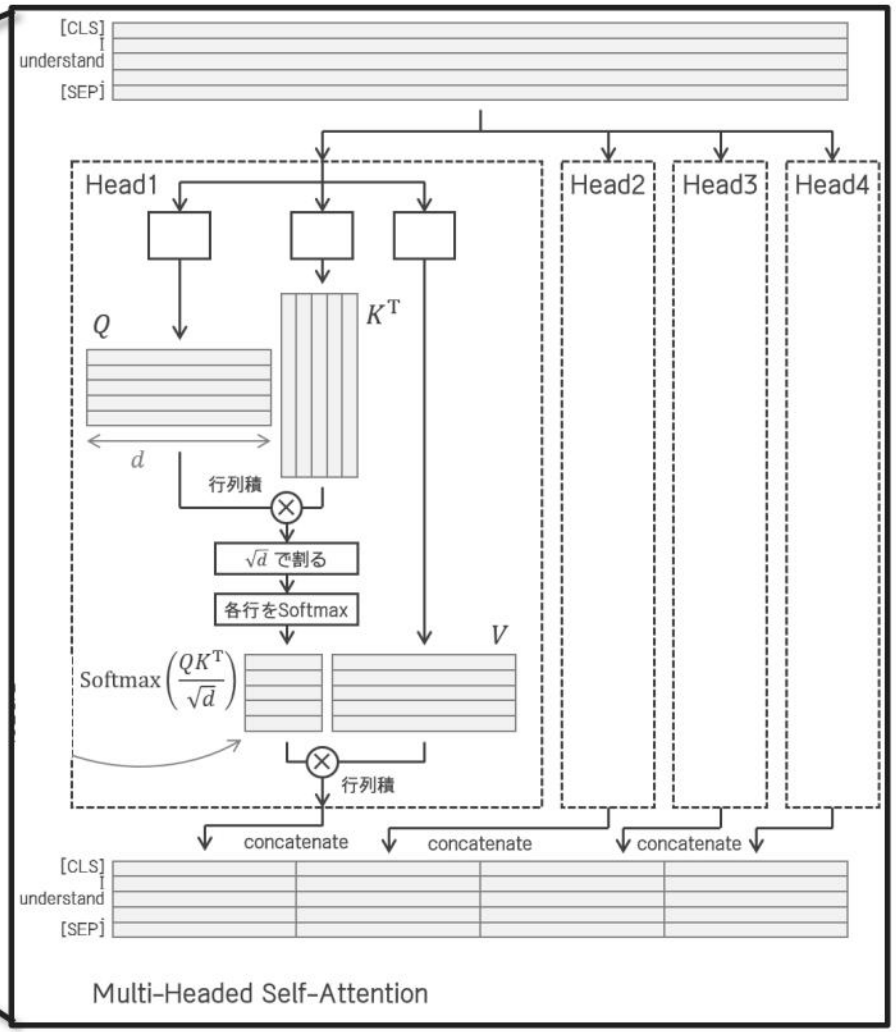
Informer

The core of Transformer models



Transformer

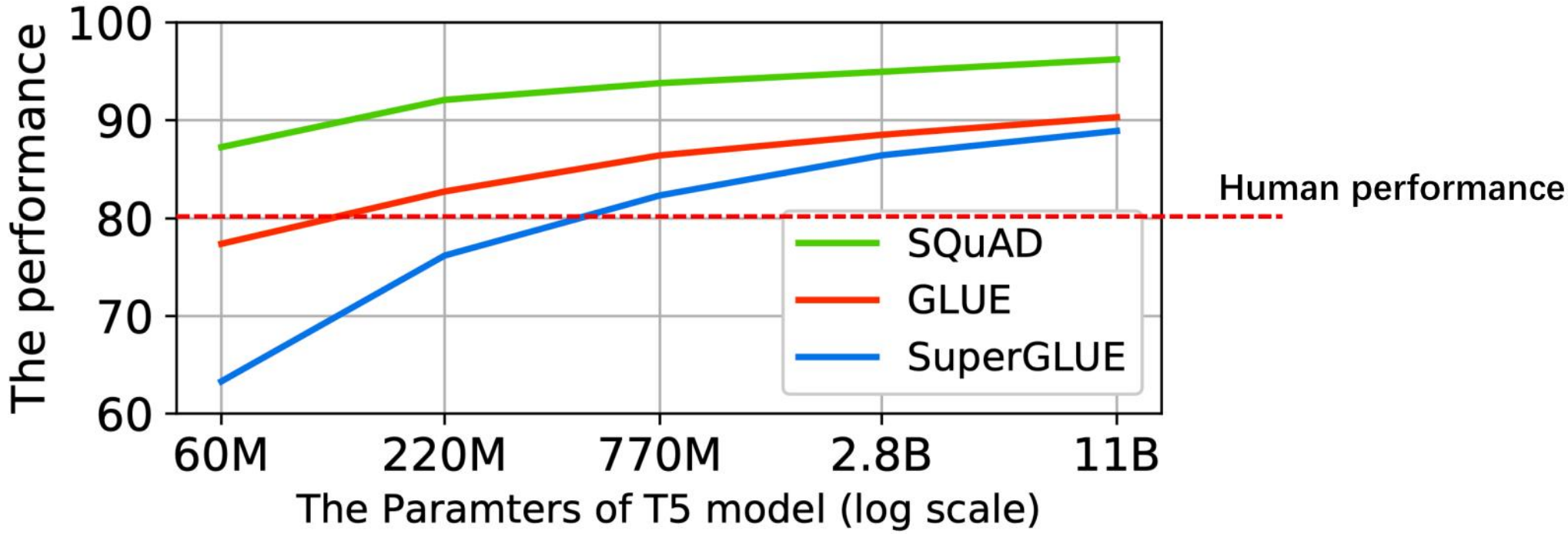
Self-attention



Multi-Headed Self-Attention

Images are acquired from the Google search engine.

Performance growth of Transformer



The triplet relationship

"Kate likes the apple"

"We are in the garden"

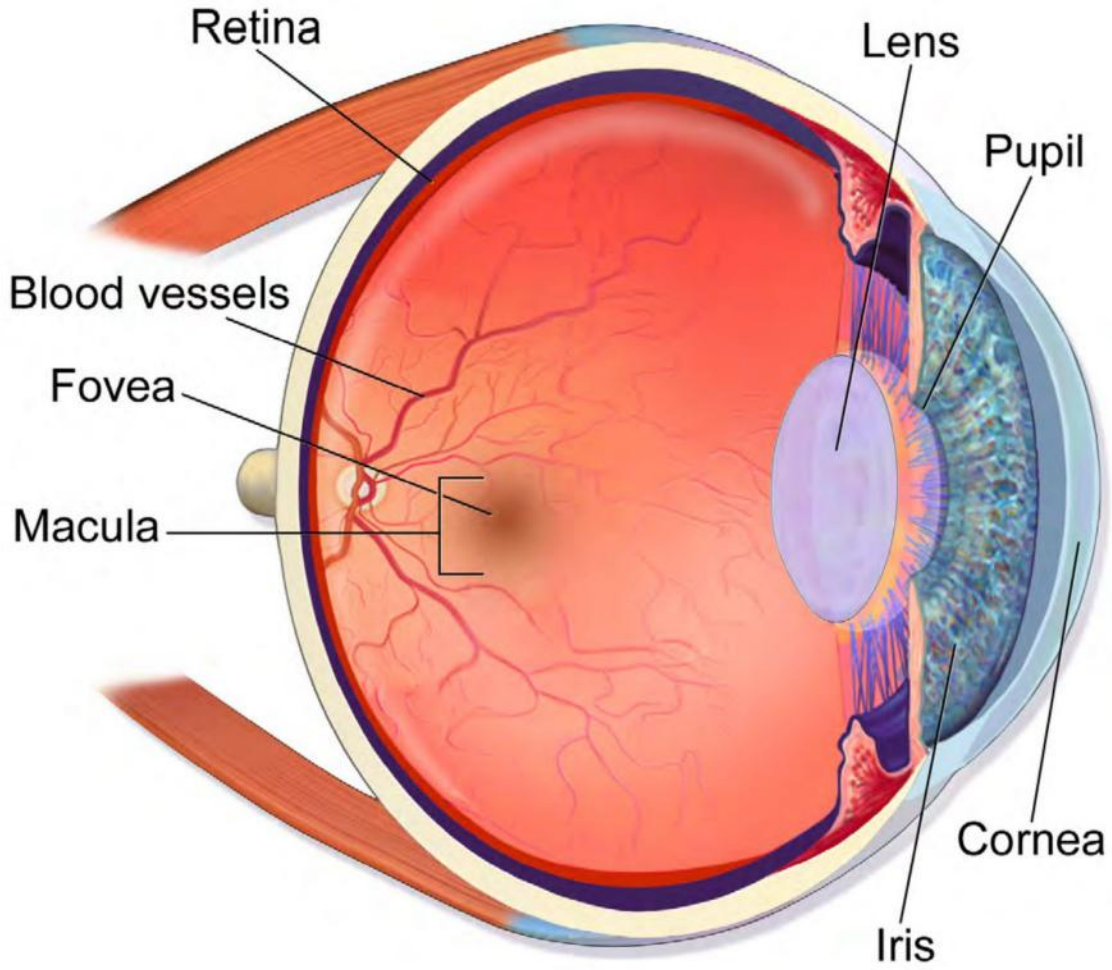
"We are in the Bestbuy"

Images are acquired from the Google search engine.

Content

- Background
- **The Triplet Attention**
- Experimental Results
- Summary & Future Work

What is the attention?



Images are acquired from the Google search engine.

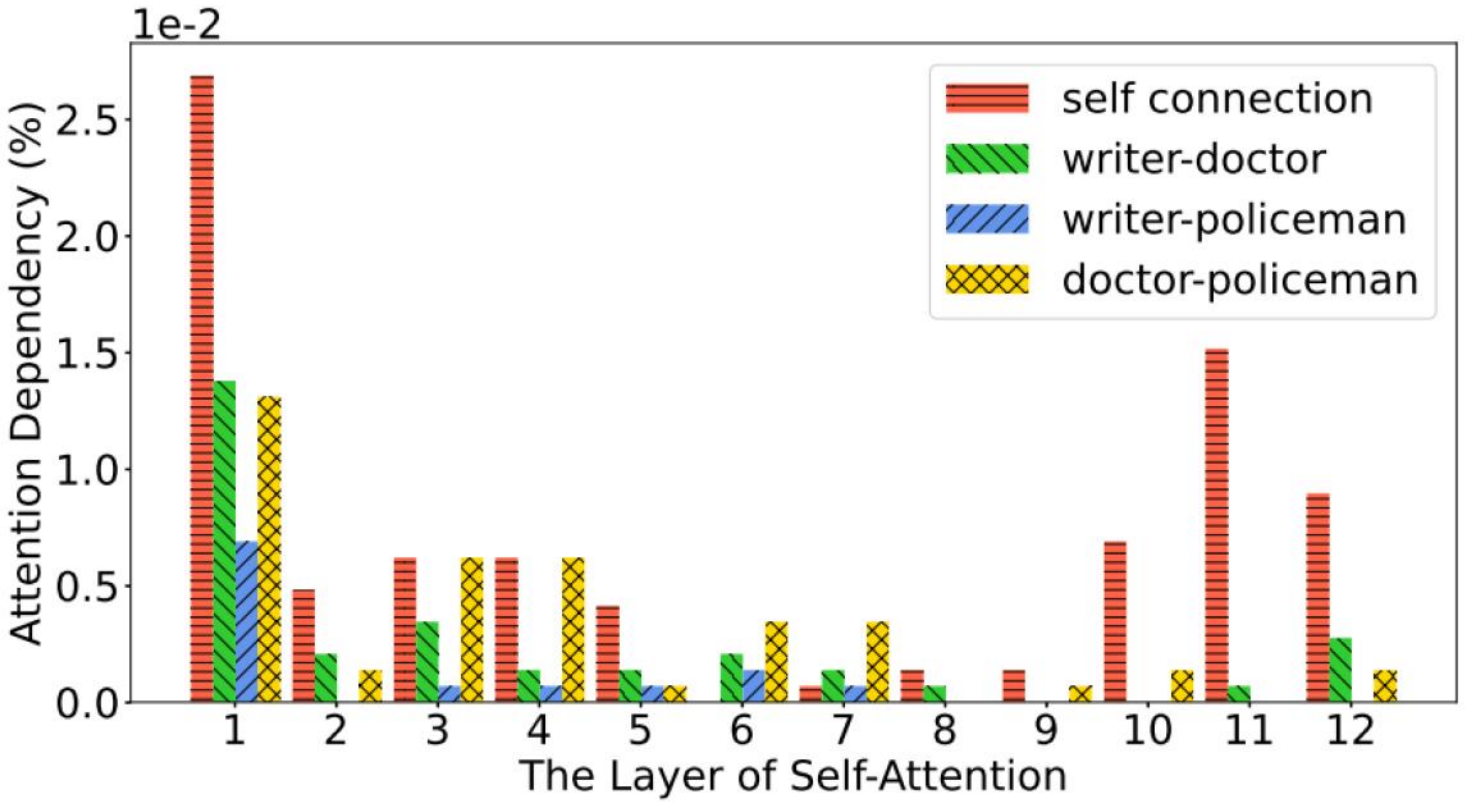
From a logical reasoning questions

"The old writer is the father of the doctor."

"The doctor is the mother of the policeman on duty at the door."

Rethinking the Canonical Self-attention

Model: BERT
with Pre-trained model



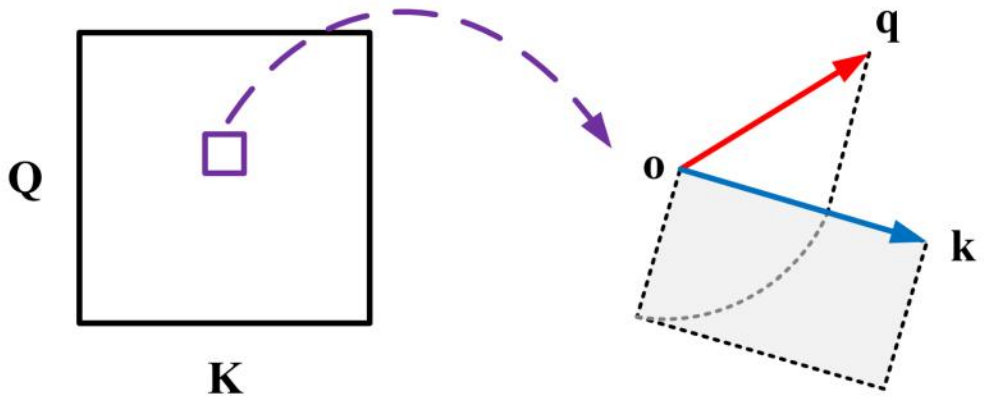
1. There are many self connections.
2. For the triplet connections between “writer-doctor-policeman”, it is much less than self connections.
3. The triplet attention decrease with the layer stacking in the BERT model.

The original formulation of Self-attention

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} \quad \Rightarrow \quad \mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{D}^{-1}\mathbf{A}\mathbf{V}$$

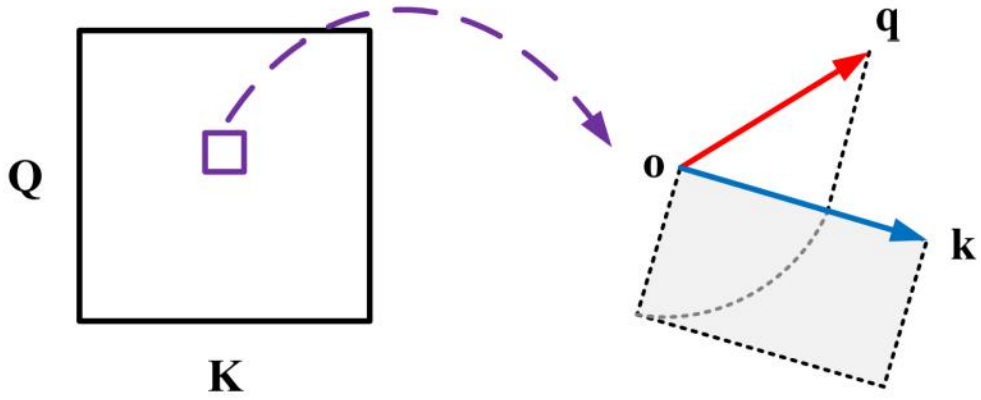
Attention Matrix: $\mathbf{A} = \exp(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})$

Normalizer: $\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1}_L)$



The grey square denotes the score of self-attention mechanism.

The original formulation of Self-attention

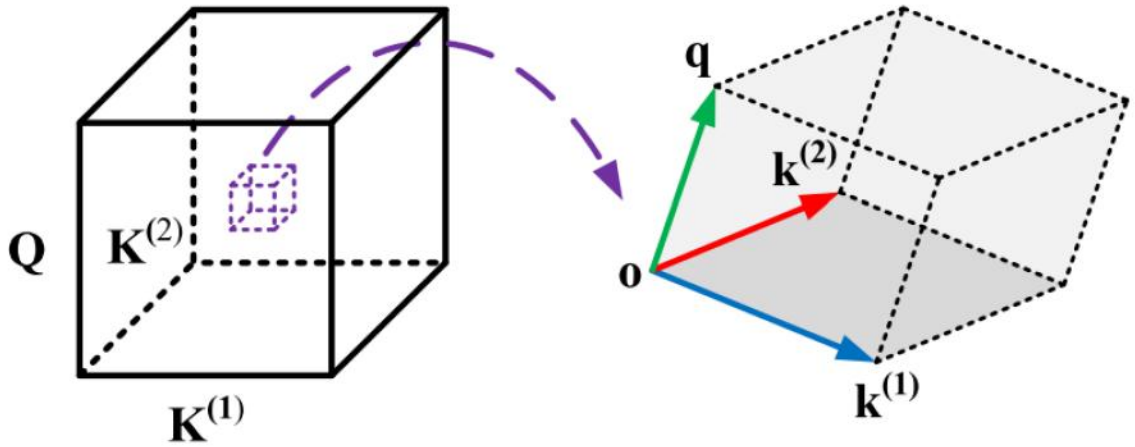


Measure the pairwise score of inputs.

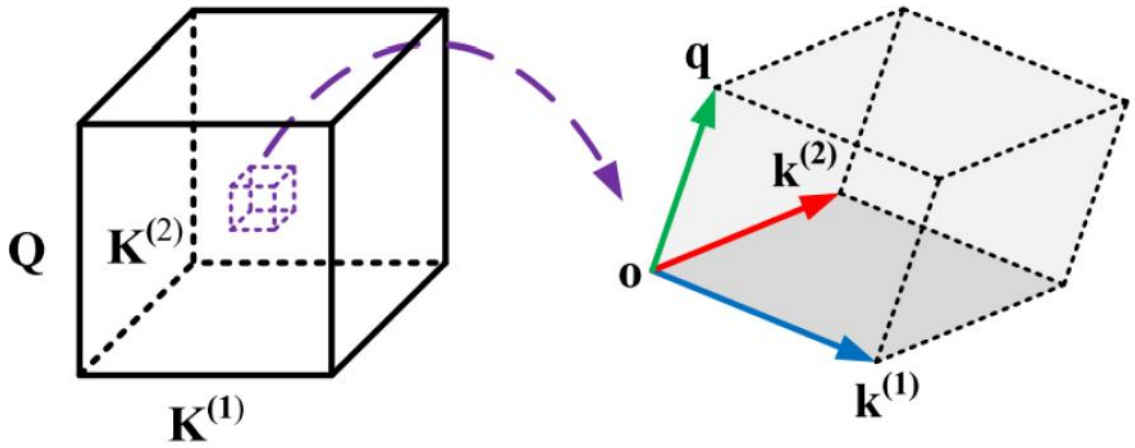
(Mutual information)

Measure the triplet score of inputs.

(Diversity)



The Triplet Attention



Add another Key, we have the normal vector as:

$$\mathbf{K}_{j\times k} = \mathbf{K}_j^{(1)} \times \mathbf{K}_k^{(2)}$$

We can define the triplet similarity as:

$$\mathrm{sim}(i,j,k) = \exp\left(\frac{T_{ijk}}{\sqrt{d_t}}\right)$$

where T_{ijk} is the Scalar Triplet Product (STP):

$$T_{ijk} = \mathbf{Q}_i \mathbf{K}_{j\times k}^{\top} = \mathbf{Q}_i (\mathbf{K}_j^{(1)} \times \mathbf{K}_k^{(2)})^{\top}$$

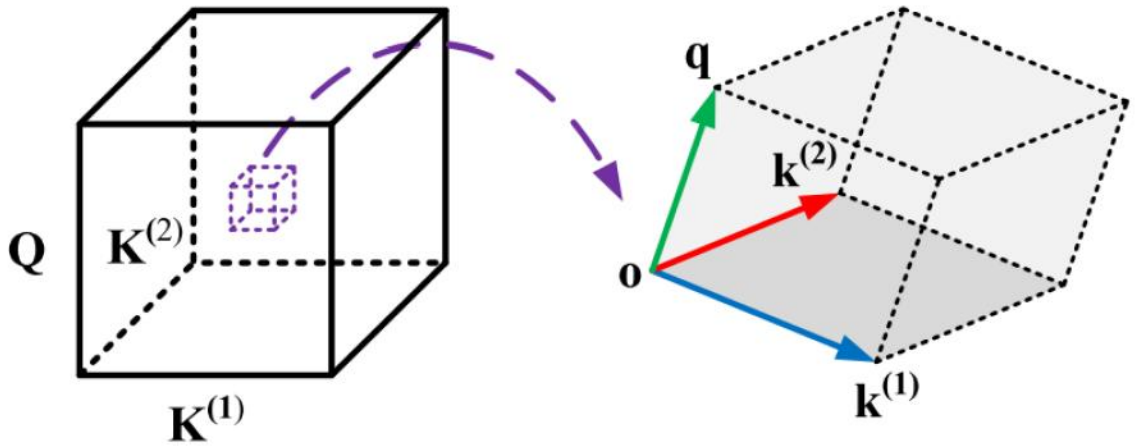
Note that STP can only be computed at 3-dim and 7-dim.

Thus, the A^3 attention is defined:

$$\mathrm{tAttention}(\mathbf{Q}, \mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \mathbf{V}) = \mathcal{D}^{-1}\mathcal{A}\mathbf{V}$$

$$\mathcal{A} = \exp\left(\frac{\mathbf{TW}}{\sqrt{d_t}}\right)$$

The efficient A^3 Attention



Thus, the A^3 attention is defined:

$$\text{tAttention}(\mathbf{Q}, \mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \mathbf{V}) = \mathcal{D}^{-1}\mathcal{A}\mathbf{V}$$

$$\mathcal{A} = \exp\left(\frac{\mathbf{TW}}{\sqrt{d_t}}\right)$$

Using the Permuted STP for fast inter-position information sharing: $\mathbf{T}_{\mathcal{P}} = \mathcal{P}^{-1}\left[\mathbf{Q}(\mathbf{K}^{(1)} \otimes \mathcal{P}[\mathbf{K}^{(2)}])^{\top}\right]$

The efficient A^3 attention is defined:

$$\text{atAttention}(\mathbf{Q}, \mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \mathbf{V}) = \mathfrak{D}^{-1}\mathfrak{A}\mathbf{V}, \quad \mathfrak{A} = \exp\left(\frac{\mathbf{T}_{\mathcal{P}}}{\sqrt{d_t}}\right)$$

$$\mathfrak{A} = \exp\left(\frac{\mathcal{P}^{-1}\left[\mathcal{P}[\mathbf{K}^{(2)}](\mathbf{Q} \otimes \mathbf{K}^{(1)})^{\top}\right]}{\sqrt{d_t}}\right) \qquad \mathfrak{A}_{ij} = \mathcal{P}^{-1}\left[\exp(\frac{\|\mathbf{K}'_i\|_2^2}{2\sqrt{d_t}}) \cdot \exp(\frac{-\|\mathbf{K}'_i - \mathbf{Q}'_j\|_2^2}{2\sqrt{d_t}}) \cdot \exp(\frac{\|\mathbf{Q}'_j\|_2^2}{2\sqrt{d_t}})\right]$$

By applying the FAVOR trick: $\mathbf{B} = \phi(\hat{\mathbf{K}})\phi(\hat{\mathbf{Q}}),\ where\ \hat{\mathbf{K}}_i = \phi(\mathbf{K}'_i),\ \hat{\mathbf{Q}}_i = \phi(\mathbf{Q}'_i), \quad \phi(\mathbf{x}) \stackrel{\text{def}}{=} \sqrt{\frac{2}{c}}\left(\cos(\omega_1^{\top}\mathbf{x} + b_1), \ldots, \cos(\omega_c^{\top}\mathbf{x} + b_c)\right)$

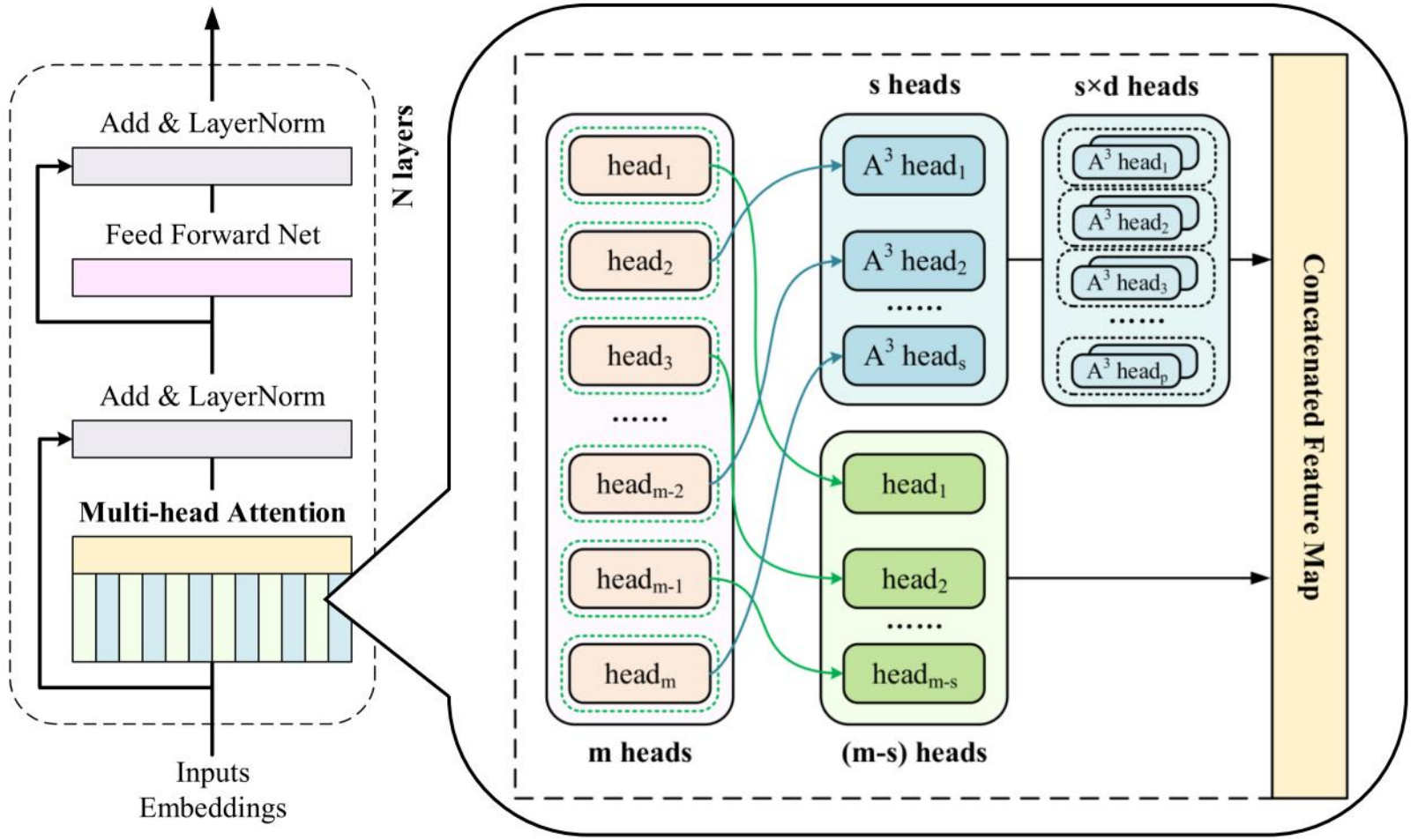
Overall architecture

The multi-head A^3 attention:

$$\text{matAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \mathbf{V}) = \text{concat}(\text{head}_1, \ldots, \text{head}_s, \ldots, \text{head}_m) \quad ,$$

$$\textit{where } \text{head}_i = \begin{cases} \text{atAttention}(\mathbf{Q}_i, \mathbf{K}_i^{(1)}, \mathbf{K}_i^{(2)}, \mathbf{V}_i), & \text{if } i \leq s \\ \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i), & \text{otherwise} \end{cases}$$

A3 is compatible with canonical self-attention and can be used interchangeably.



Content

- Background
- The Triplet Attention
- **Experimental Results**
- Summary & Future Work

Experiment settings

Datasets

- General Language Understanding Evaluation (GLUE)
- Stanford Question Answering Dataset (SQuAD v1.1)

Baselines

GLUE: ELMo, DistilBERT, BERT*base*

SQuAD v1.1: BiDAF-ELMo, R.M. Reader, DistilBERT, BERT*base*

Metrics

GLUE: The Matthews Correlation Coefficient for CoLA, Pearson Correlation Coefficient for STS-B, Accuracy for others

SQuAD v1.1: Exact Match (EM) and F1 score

Platform

All methods are run on one single Nvidia V100 GPU.

(1) GLUE Experiment Results

Model	CoLA 8.5k	MRPC 3.5k	RTE 2.5k	STS-B 5.7k	QNLI 108k	QQP 363k	SST-2 67k	WNLI 0.64k	MNLI 392k	Average -
ELMo	44.1	76.6	53.4	70.4	71.1	86.2	91.5	56.3	68.6	68.7
DistilBERT	51.3	87.5	59.9	86.9	**89.2**	88.5	91.3	56.3	82.2	77.0
DistilBERT-A^3 (ours)	**55.1**	**87.9**	**67.8**	**87.5**	88.9	**90.3**	**91.5**	**56.3**	**83.8**	**78.8**
DistilBERT-A^3-FAVOR (ours)	**55.1**	87.8	65.6	87.1	88.3	89.8	91.1	56.3	83.6	78.3
$BERT_{base}$	56.3	88.6	69.3	89.0	**91.9**	89.6	92.7	53.5	86.6	79.7
BERT-A^3 (ours)	**62.9**	**90.8**	**74.4**	**90.3**	91.1	**91.0**	**93.2**	**56.3**	**87.6**	**81.9**
BERT-A^3-FAVOR (ours)	61.8	89.8	72.2	90.1	90.8	90.8	92.7	56.3	87.4	81.3

A^3 mechanism

- $\boldsymbol{A^3}$ enhance the complementary dependency of self-attention mechanism and gain more competitive scores.
- DistilBERT-A^3 and BERT-A^3 outperform DistilBERT and $BERT_{base}$ respectively on 8 tasks.
- DistilBERT-A^3 / BERT-A^3 achieves 7.4%/11.7% score rising on CoLA and 13.2%/6.8% on RTE dataset.
- Triplet connections help the model capture weak dependency with limited data.

(2) SQuAD Experiment Results

Model	SQuAD	
	EM	F1
BiDAF-ELMo	-	85.6
R.M. Reader	81.2	87.9
DistilBERT	77.7	85.8
DistilBERT-A^3 (ours)	78.5	87.1
DistilBERT-A^3-FAVOR (ours)	78.9	87.8
$BERT_{base}$	81.2	88.5
BERT-A^3 (ours)	**81.8**	**89.3**
BERT-A^3-FAVOR (ours)	81.6	88.9

A^3 mechanism

DistilBERT-A^3 and BERT-A^3 achieve better performance in both EM and F1 scores.

(3) Ablation study: Layer Deployment

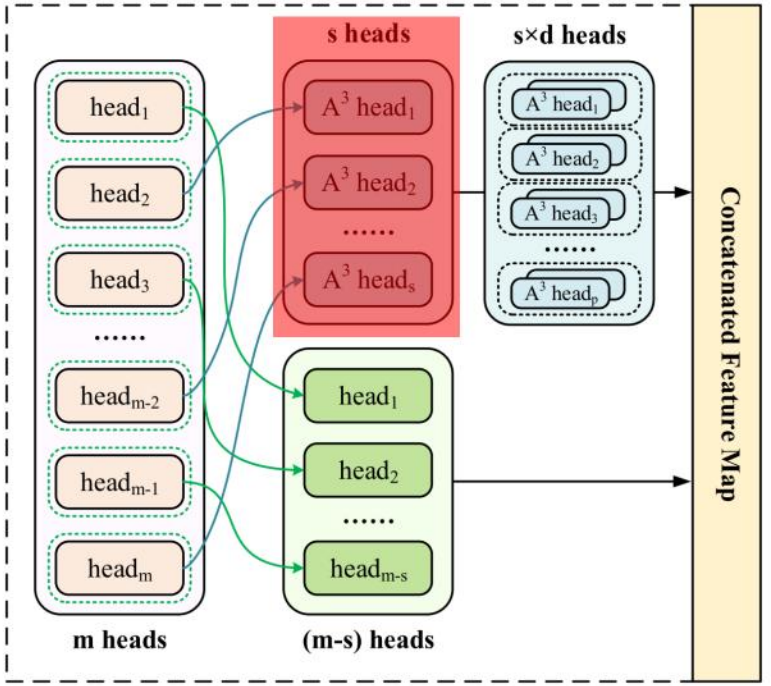
Model	CoLA	MRPC	RTE	STS-B
BERT-A$^3_{l[1-3]}$	**60.8**	**90.1**	**71.8**	89.9
BERT-A$^3_{l[4-6]}$	59.4	87.8	70.4	89.7
BERT-A$^3_{l[7-9]}$	58.6	88.7	67.9	**90.0**
BERT-A$^3_{l[10-12]}$	59.6	88.9	67.1	89.7
BERT-A$^3_{l[1-6]}$	58.9	88.8	68.6	89.7
BERT-A$^3_{l[7-12]}$	57.9	89.4	70.1	89.7
BERT-A$^3_{l[1-12]}$	56.3	87.6	69.0	89.4

Model	CoLA	MRPC	RTE	STS-B
BERT-A$^3_{l\{1,3\}}$	**60.8**	**89.8**	**74.4**	90.0
BERT-A$^3_{l\{1,3,5\}}$	60.4	89.2	73.3	**90.2**
BERT-A$^3_{l\{2,4\}}$	58.9	**89.8**	73.6	89.8
BERT-A$^3_{l\{2,4,6\}}$	60.7	89.4	71.1	90.0
BERT-A$^3_{l\{1,12\}}$	58.5	89.3	73.6	89.5
BERT-A$^3_{l\{1,3,10\}}$	60.3	89.6	71.8	90.0
BERT-A$^3_{l\{1,3,10,12\}}$	59.9	89.8	72.6	**90.2**
BERT-A$^3_{l\{2,11\}}$	59.1	89.6	72.2	89.6
BERT-A$^3_{l\{2,4,11\}}$	**61.1**	**90.0**	**74.0**	89.7
BERT-A$^3_{l\{2,4,9,11\}}$	60.4	89.3	66.8	89.7

- Using A^3 heads in lower consecutive layers leads to better scores.
- Adding A^3 layer in higher layers receives reduced performance improvement, but still better than BERT*base*.
- Employing A^3 mechanism in the first layer and third layer achieve more competitive scores.

(3) Ablation study: Heads Configuration

Model	CoLA	MRPC	RTE	STS-B
BERT-$A^3{}_{h3}$	**60.8**	**90.1**	**71.8**	**89.9**
BERT-$A^3{}_{h6}$	**60.8**	89.6	70.1	89.5
BERT-$A^3{}_{h9}$	57.6	88.7	66.8	88.3
BERT-$A^3{}_{h12}$	53.6	85.7	57.3	88.1



- Increasingly adding A^3 heads do not constantly improve the metrics.
- Interferes between triplet correlations and pairwise correlations: A^3 mechanism should not dominate the attention block, but it could be used as a supplement for better diversity.

(3) Ablation study: Permutation Grouping

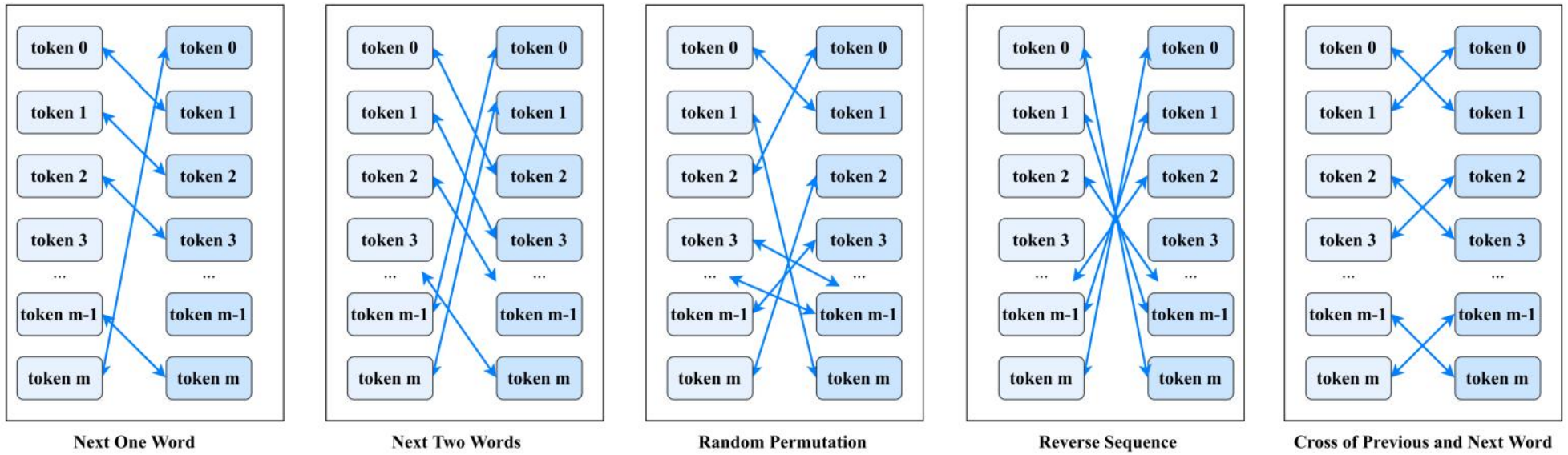
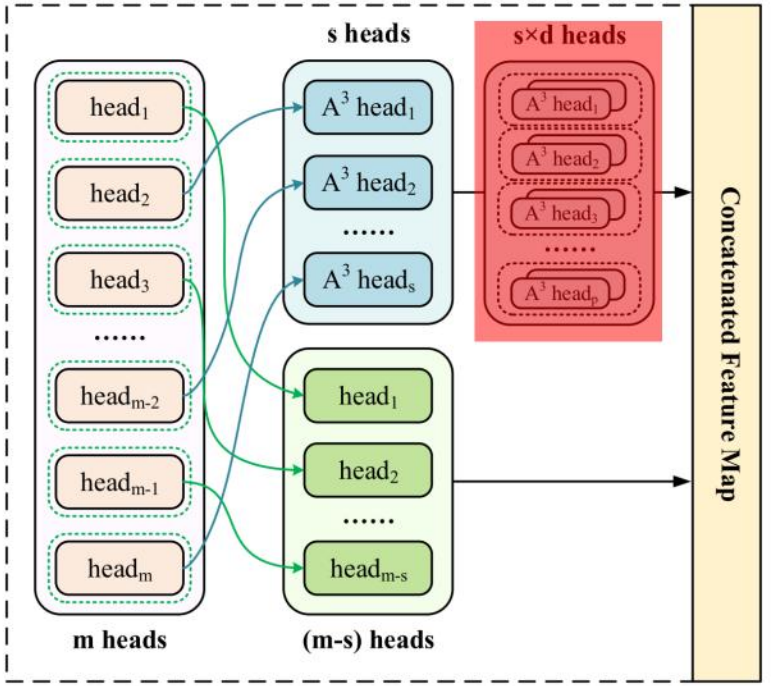


Figure 9: The illustration of five different permutation strategies.

- P1: “next one word”: shift the sequence forward by one word, and fill the last word with the first word.
- P2: “next two words”: shift the sequence forward by two words, and fill the last two words with the first two words.
- P3: “random permutation”: shuffle the sequence randomly.
- P4: “reverse sequence”: reverse the sequence.
- P5: “cross of previous and next word”: swap the values of odd index and even index from 0 and 1 in turn.

(3) Ablation study: Permutation Grouping

Model	CoLA	MRPC	RTE	STS-B
BERT-$A^3{}_{P\{1\}}$	**61.6**	88.9	70.4	89.6
BERT-$A^3{}_{P\{2\}}$	60.2	89.9	**71.1**	89.7
BERT-$A^3{}_{P\{3\}}$	60.8	89.3	70.0	89.4
BERT-$A^3{}_{P\{4\}}$	58.6	**90.0**	69.7	89.4
BERT-$A^3{}_{P\{5\}}$	58.6	88.9	70.8	**89.8**
BERT-$A^3{}_{P\{1,2\}}$	61.6	88.5	70.8	89.6
BERT-$A^3{}_{P\{1,3\}}$	59.8	**90.8**	**71.8**	**90.0**
BERT-$A^3{}_{P\{1,4\}}$	60.6	88.9	70.4	89.3
BERT-$A^3{}_{P\{1,5\}}$	**61.8**	88.8	71.5	89.6
BERT-$A^3{}_{P\{2,3\}}$	60.4	89.8	71.1	**90.0**
BERT-$A^3{}_{P\{1,2,3\}}$	60.8	89.6	71.8	90.2
BERT-$A^3{}_{P\{2,3,4\}}$	59.9	89.5	**72.6**	90.0
BERT-$A^3{}_{P\{1,2,3,4\}}$	61.8	89.9	71.4	90.2
BERT-$A^3{}_{P\{2,3,4,5\}}$	**62.3**	**90.8**	70.8	**90.3**



- The "random permutation" performs well because it increases the probability of cross product between two distant dissimilar vectors, thereby discovering more dependency.
- Increasing the variety of Permuted STP methods can obtain more significant dissimilar information and enhance the diversity of attention mechanism.

(4) Case Study: Layer Stacking Degradation

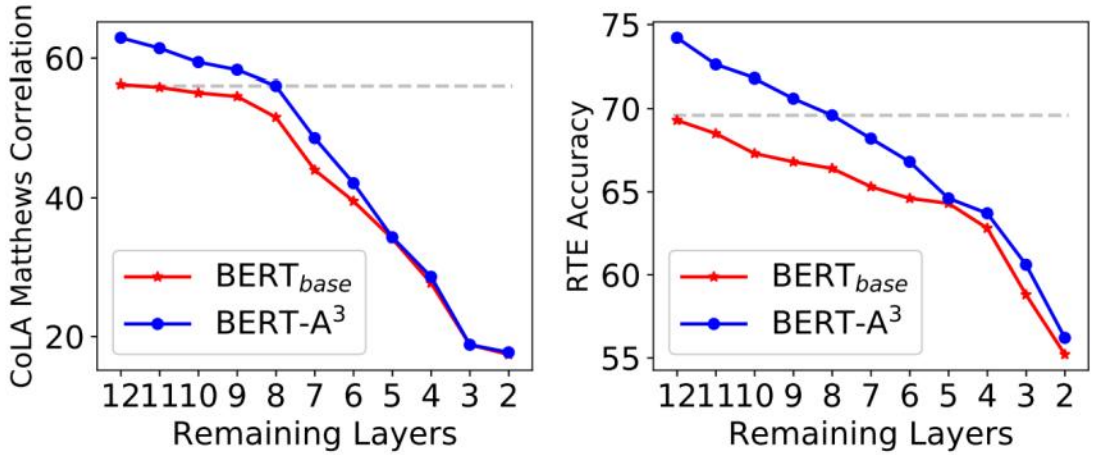


Figure 5: The performance decreases when the layer number of Transformers degrades from 12 to 2.

- BERT-A^3 consistently outperforms BERT*base* with fewer layers.
- BERT-A^3 with only eight layers outperforms BERT*base* with complete 12 layers on both datasets.
- We can reduce the number of the parameters by 25.6% and the computation time by 28.9% by using A^3 mechanism while still achieving similar or better results.

(4) Case Study: Applying the FAVOR Trick

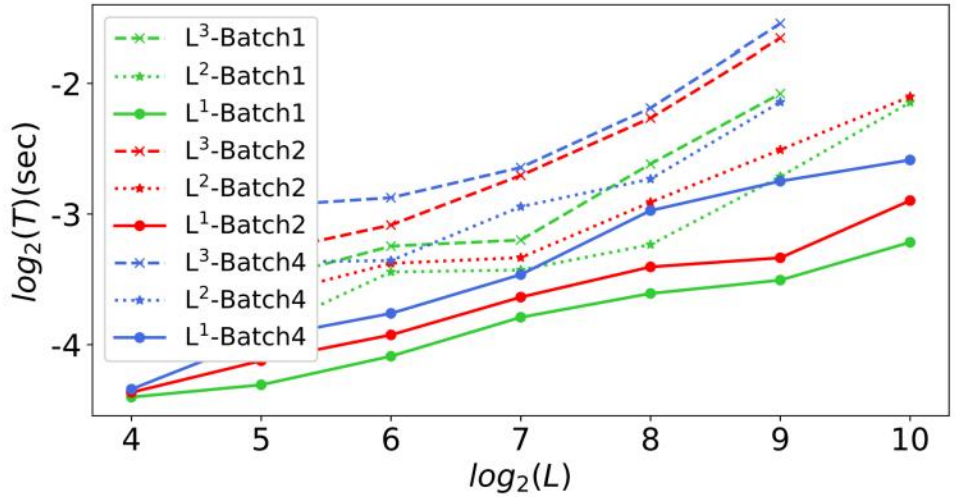
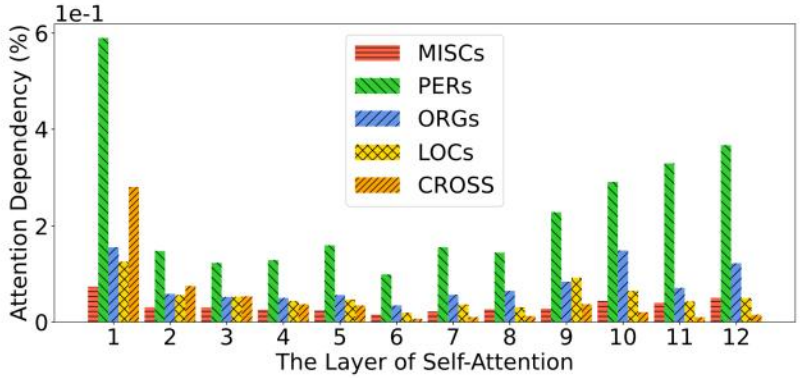


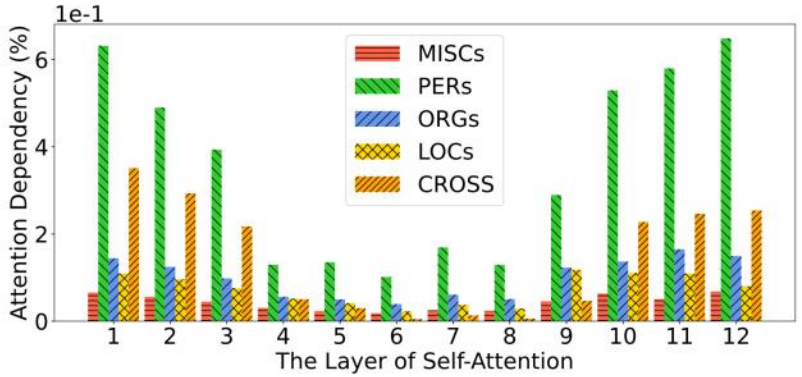
Figure 6: Comparison of BERT-A^3-L^3, BERT-A^3, BERT-A^3-FAVOR in terms of forward and backward pass speed.

- BERT-A^3-FAVOR reduce the computation cost, and BERT-A^3-FAVOR model allows extensive batch training and lower computation time, which contributes to total train time reduction.

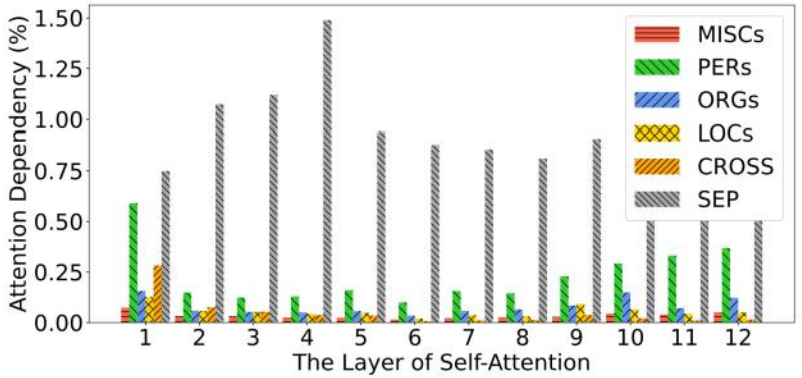
(4) Case Study: NER Visualization



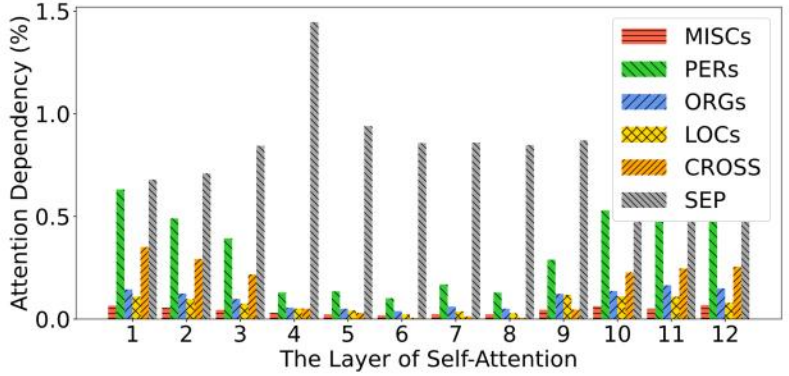
(a) The NER visualization on the BERT$_{base}$ model.



(b) The NER visualization on the BERT-A$^3{}_{l\{1,2,3,10,11,12\}}$ model.



(a) The NER visualization on the BERT$_{base}$ model.



(b) The NER visualization on the BERT-A$^3{}_{l\{1,2,3,10,11,12\}}$ model.

- BERT-A^3 significantly increases the number of 'CROSS' connections and other connections at the layers {1,2,3,10,11,12}, where we apply the A^3 mechanism.

Content

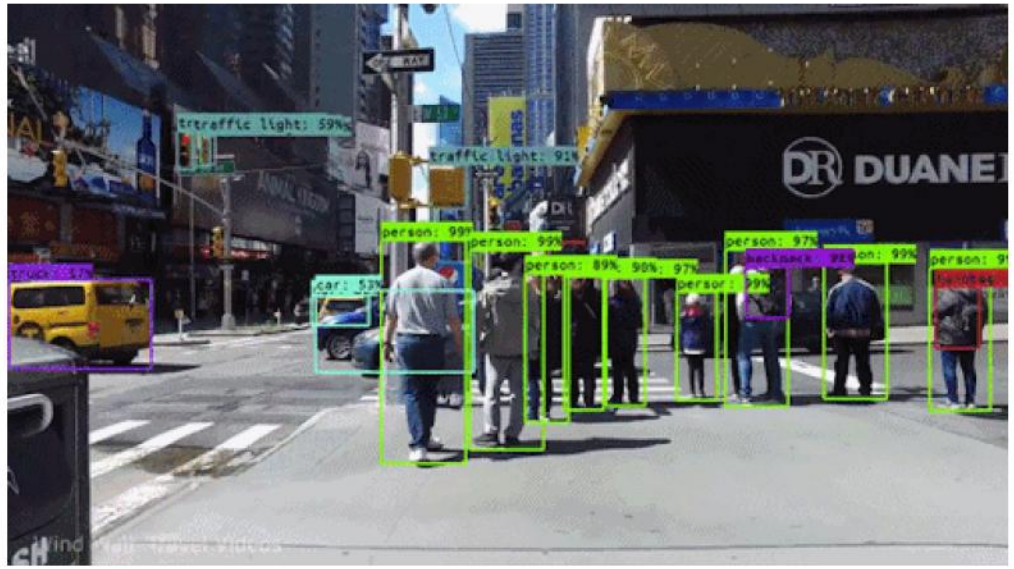
- Background
- The Triplet Attention
- Experimental Results
- **Summary & Future Work**

Summary

- Slightly breaking the similarity assumption in dot-product self-attention computation
- Triplet Attention mechanism allows for attention on dissimilarity pairs and it contributes to building the high-level dependency
- Triplet Attention mechanism can be deployed with canonical self-attention through proper configurations.
- A^3-FAVOR: efficient variants for long inputs on large-scale models.
- Experimental results on two benchmarks demonstrate that A^3 significantly outperforms the baselines and it shows the benefits of introducing triplet attention into Transformers.

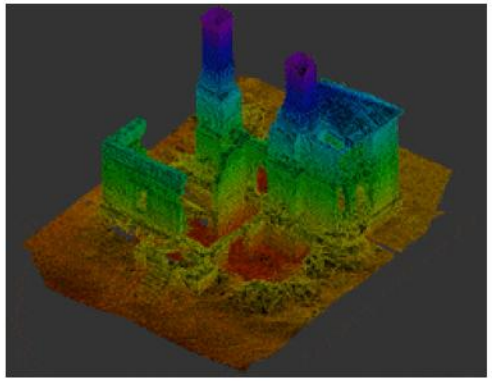
Future Work

Music Style



Video Analysis

3D Point



Thank you!

Presented by: Haoyi Zhou (www.zhouhaoyi.com)
Email: zhouhy@act.buaa.edu.cn

Discussion (online)

BEIHANG
UNIVERSITY