Enhanced BERT for Natural Language Inference and Sentence Classification

楚天翔

chutianxiang@gmail.com

Background

- Natural Language Inference

Natural language inference (NLI) is the task of determining the inferential relationship between two or more sentences.

A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

Background

- Sentence-embedding method vs. Cross-sentence Method

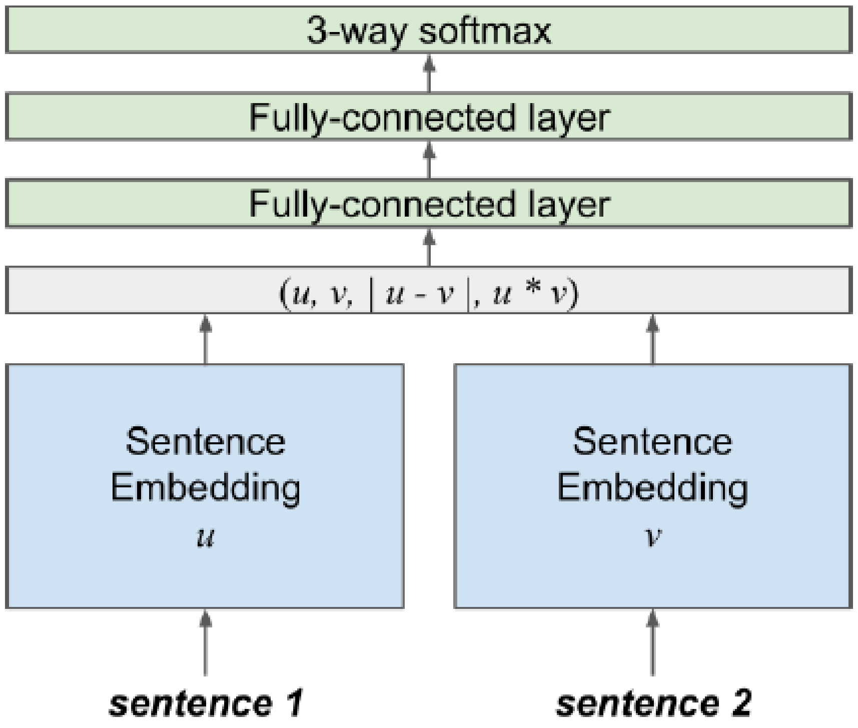


Figure from *Natural Language Inference with Hierarchical BiLSTM Max Pooling Architecture*

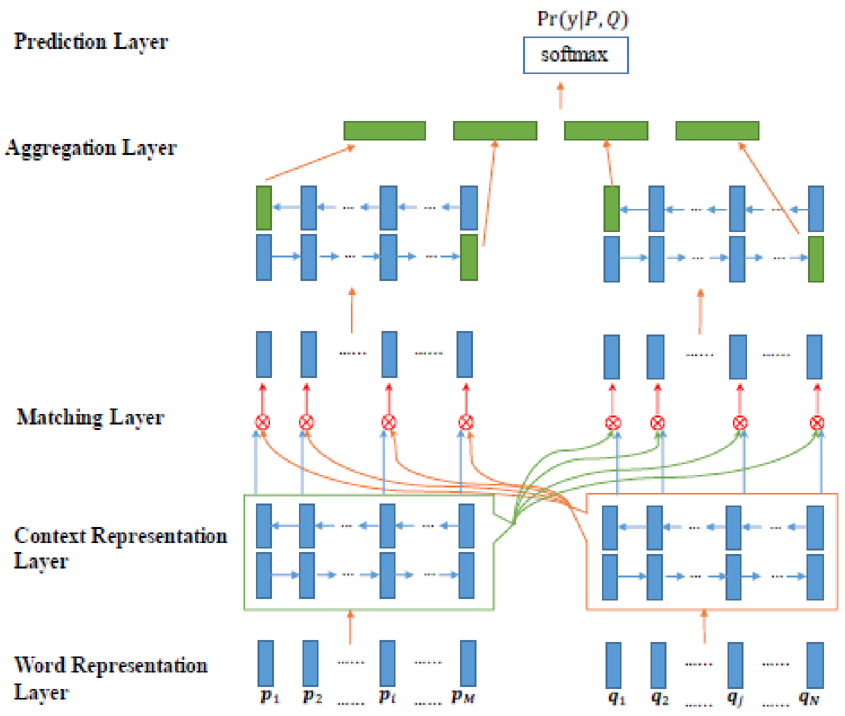


Figure from *Bilateral Multi-Perspective Matching for Natural Language Sentences*

Background

- BERT

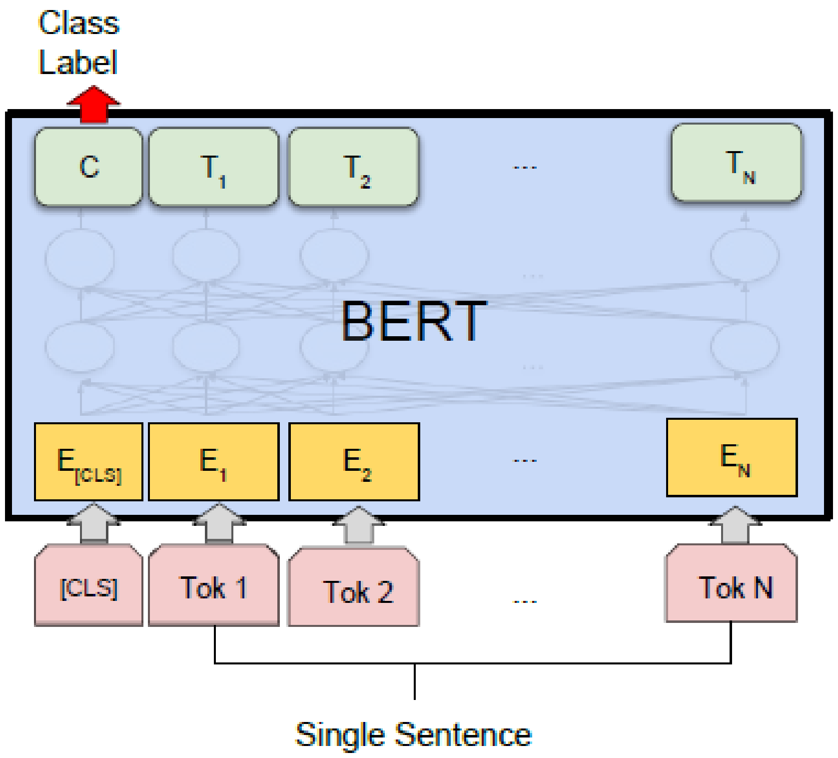


Figure from *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

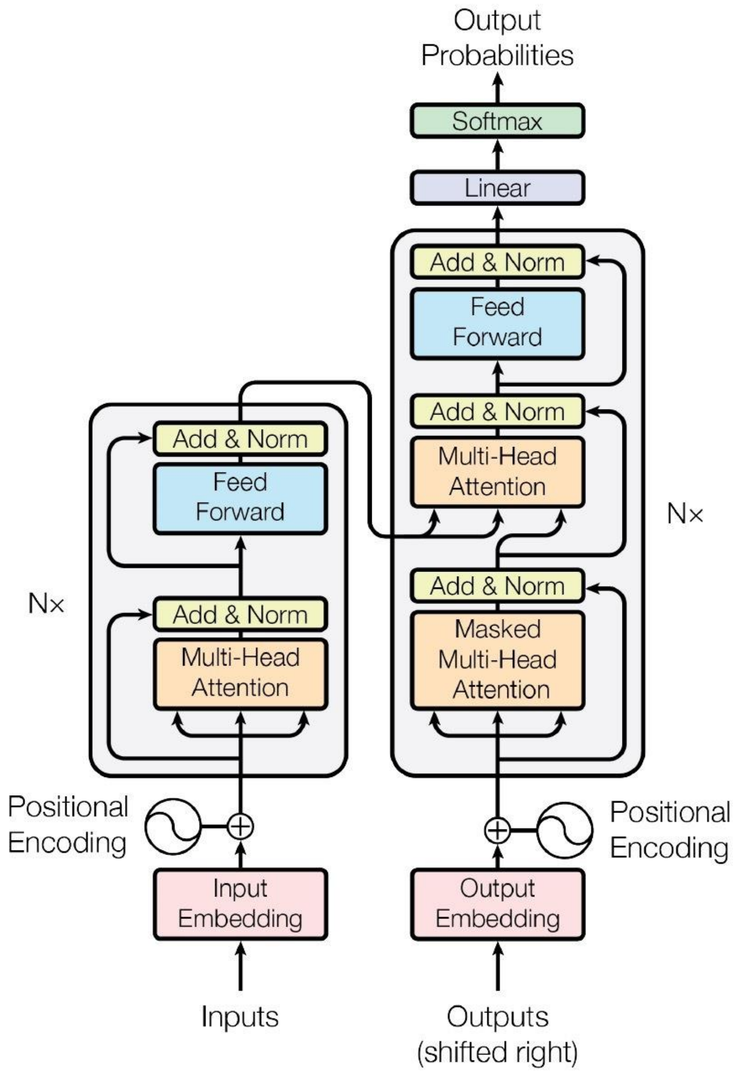
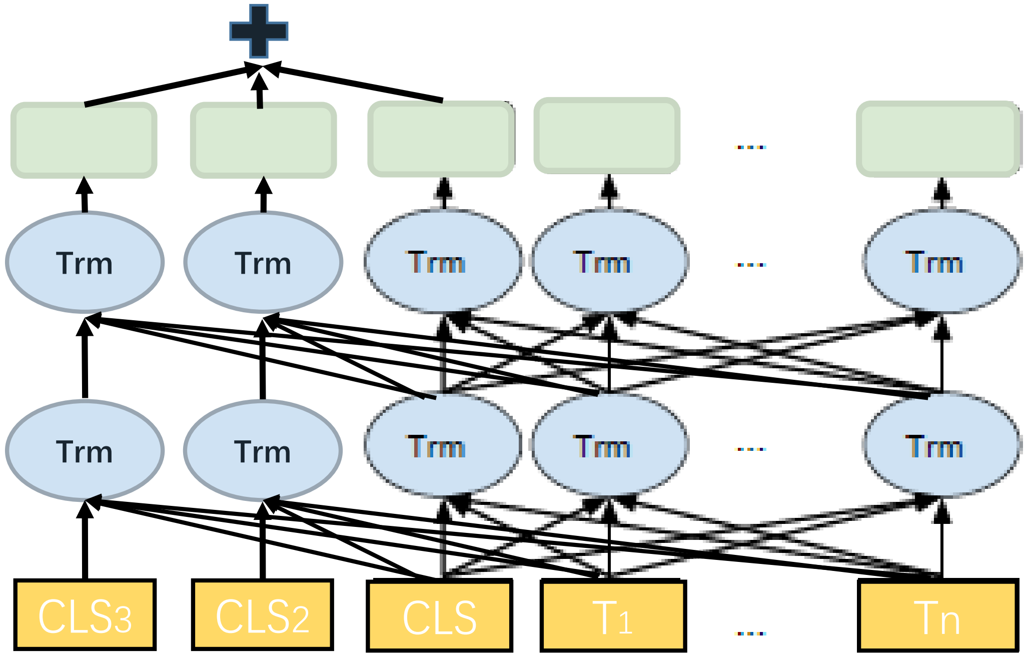


Figure from *Attention is All You Need.*

Method

- As baseline, I simply concatenate the [CLS] token representation from the top four hidden layers as the sentence encoding and feed into the two-layer MLP.
- I propose to extend the [CLS] pooling method to a multi-head way (“multi-head over multi-head”).



Method

- I also tried replacing the [CLS] pooling with generalized pooling proposed in *Enhancing Sentence Embedding with Generalized Pooling.*

$$\mathbf{A} = \mathrm{softmax}(\mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \mathbf{H}^{\mathrm{T}} + b_1) + b_2)^{\mathrm{T}}$$

- Currently I only use single-head generalized pooling with no penalization term, more experiments will be carried on later.

Method

- Inspired by *BERT on STILTS*, pretraining with an intermediate task may help downstream tasks.
- After finetuning on the SNLI dataset, I further finetune it on SST and CoLA datasets to see if it can bring any improvement.

Result

- SNLI

Method	Dev set acc (%)	Test set acc (%)
600D Hierarchical BiLSTM with Max Pooling	-	86.6
600D BiLSTM with generalized pooling	-	86.6
512D Dynamic Meta-Embeddings	-	86.7
2400D Multiple-Dynamic Self-Attention Model	-	87.4
Baseline	87.9	87.4
Multi-CLS	**88.3**	**87.7**
Generalized Pooling	88.2	87.6

Result

- GLUE
- I also test the proposed multi CLS method in GLUE too.
- Currently experiments are only performed on 5 relative small datasets(CoLA, MRPC,RTE, SST-2, STS-B). The following scores are on dev set.

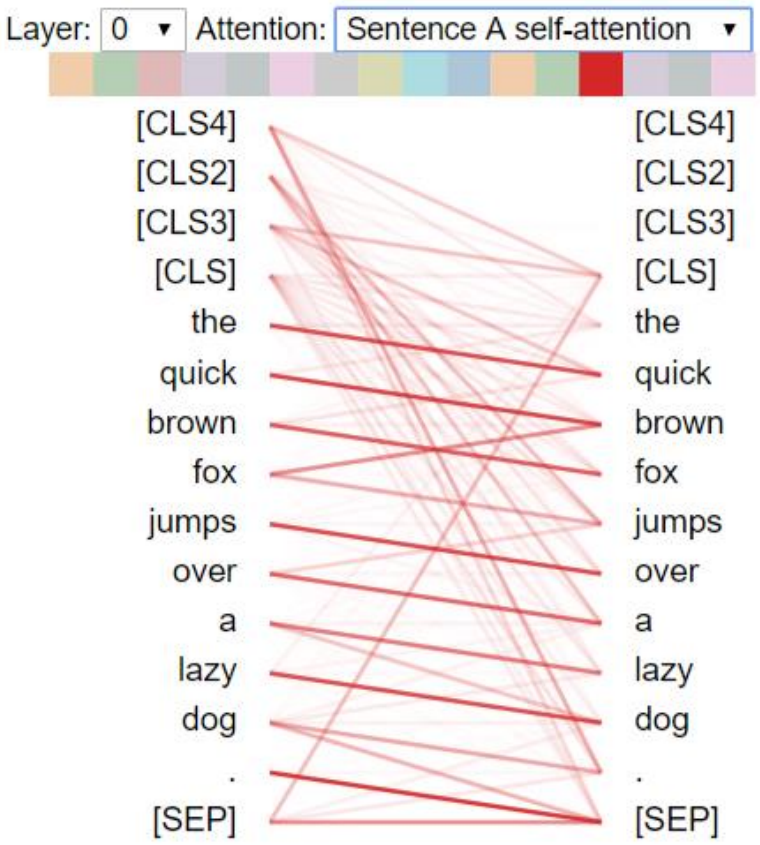
Dataset	Single CLS	Multi-CLS
CoLA	63.3	**65.8**
SST-2	94.2	**94.4**
MRPC	89.0/92.2	**89.2/92.3**
RTE	74.0	**75.1**
STS-B	90.2/90.0	**90.5/90.3**

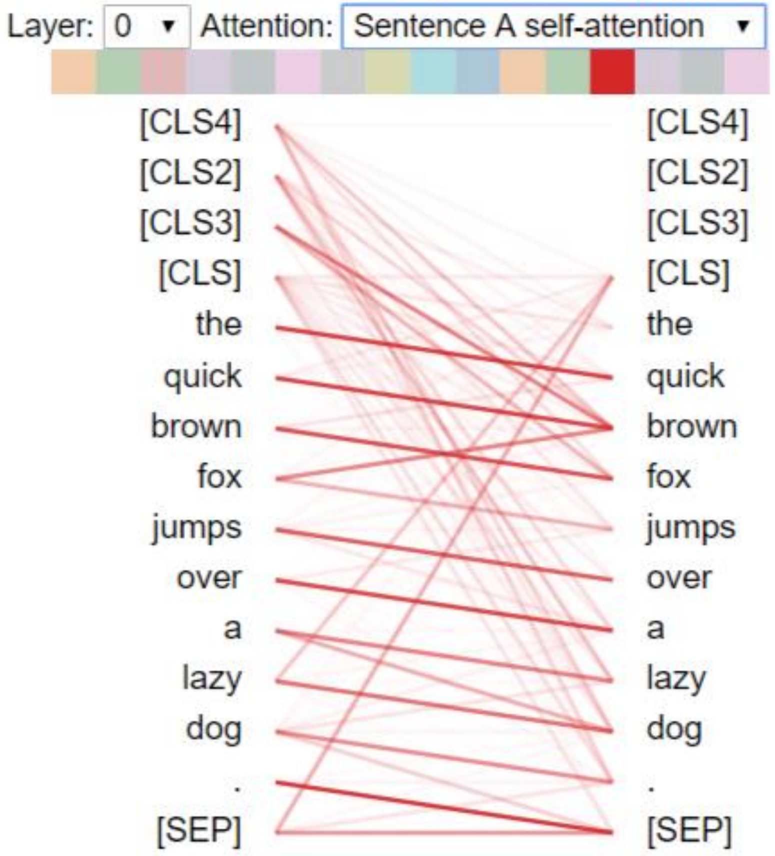
Result

- SNLI as Intermediate Task.
- Direct finetune consistently yields better results.

Datasets	Direct Finetune	SNLI + Finetune
SST	**94.2**	93.4
CoLA	**63.3**	61.6

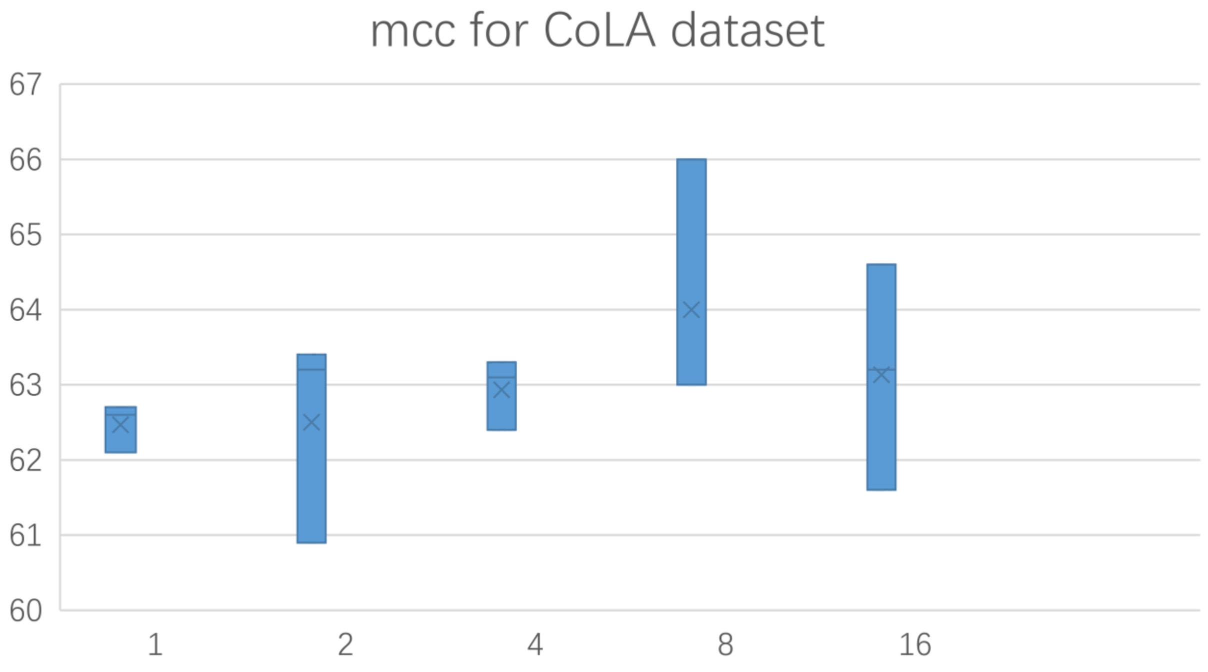
Analysis

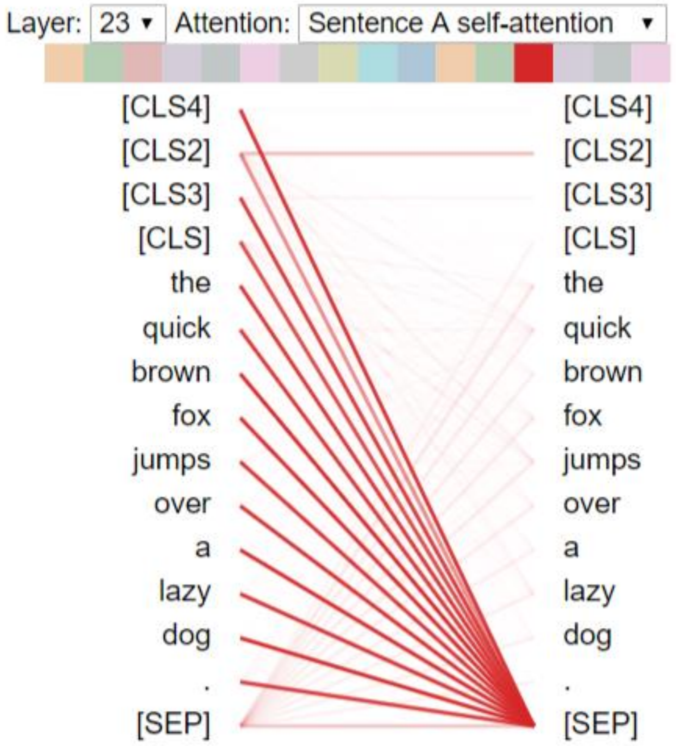

Layer: 0
Attention: Sentence A self-attention
[CLS4]
[CLS2]
[CLS3]
[CLS]
the
quick
brown
fox
jumps
over
a
lazy
dog
.
[SEP]


Layer: 0
Attention: Sentence A self-attention
[CLS4]
[CLS2]
[CLS3]
[CLS]
the
quick
brown
fox
jumps
over
a
lazy
dog
.
[SEP]

Analysis

- Different Layers





Thank You