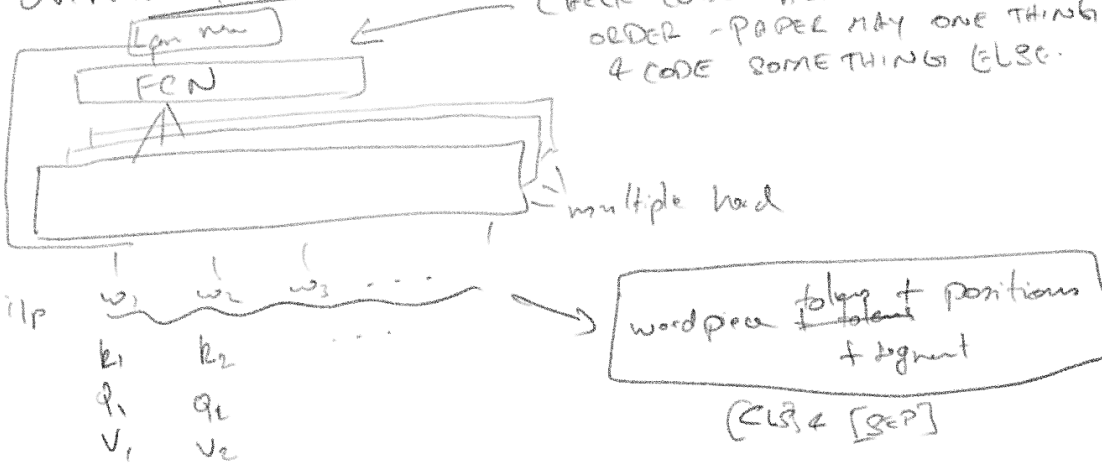
A primer in BERTology

Feb 2020 - Umass Lowell

Abs
Transformer used a lot - but do not understand inner working
This paper describes what is known to date
Propose modifications for further research

① Limits hypothesis driven development blog of lack of understanding of inner workings
- Unlike CNN transformer have little cognitive motivation
- Size hindrance to even ablation studies

② Overview of BERT architecture



2 stages
a) Pretraining b) Fine-tuning

2 tasks
MLM NSP (next sent prediction)

2 versions
a) Base b) Large

③ BERT EMBEDDINGS

word2vec & Glove are static, BERT repr is contextualized.

④ BERT embedding form clear clusters corresponding to

word order

(b) Late layers in BERT produce more context specific repr [How did they measure this?]

(4) What knowledge does BERT have?

Popular approaches

- fill-in-the-gap probes
- Analysis of Attn weights
- probing classifiers

(4.1) Syntactic knowledge

a) BERT representations are hierarchical rather than linear. In addition to apart from word order information, there is also something like a syntactic tree - Lin 2019 (Getting inside BERTs linguistic knowledge)

b) BERT embeddings also encode POS, syntactic chunks and roles [what do you learn from context? - 2019]

c) Hewitt & manning learned transformation matrices that recovers the dependency

* d) Jawahar 2019 - Approximate BERT repr with Tensor Product Decomposition Networks.

* e) Predictions were not altered with shuffled word order, truncated sentences, removed subjects & objects

d) BERT encodes syntactic structure, but does not rely on that knowledge.

(4.2) Semantic knowledge

(a) tip a robin < tip a chef < tip a waiter

incorrect fillers for semantic roles that are semantically related rather than not related

[How was this preference measured?]

(b) BERT struggles with representations of numbers - probes? whi

due to word piece tokenizer [what are word pieces ...
not straight away start with w2Vec or GloVe?]
[Also, how are numbers encoded in w2Vec?]

4.3 World knowledge

a) "Dante was born in [MASK]."

knowledge induction by filling in the blanks.

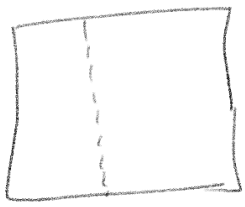
b) common sense - [BERT knows that people can walk into houses, and houses are big, but it doesn't know that houses are bigger than people — [How did they evaluate this? Does BERT "know" what is "big" / "small"? If it can learn it, then do you say it has common sense?

5) Localizing Linguistic knowledge.

5.1) Self-attention Heads

a - Estimating dark secrets of BERT. [Look this - they've classified attention types using CNNs].

Much of the model encode [SEP] and [CLS] and predictions for vertical attn



← vertical pattern in attn matrix

this redundancy must be related to overparametrization

[Did I notice this? No, I don't think so, my patterns were very different, but they all had a vertical "look" to it].

Periods, commas, are as frequent as [CLS] & [SEP], so the model learns to rely on them a lot more.

[SEP] gets increased attn starting layer 5, but its importance in prediction drops [How was this importance measured?]

5.2 BERT Layers

(a) Decrease in linear word order information around layers. Accompanied by increased knowledge of Hierarchical Sentence Structure (Lin 2019)

(b) Syntactic information is the most prominent in the middle BERT layers. Hewitt & Manning: 6-9 layers reconstructed the middle BERT layers.

Goldberg 2019: Subject-verb agreement

(c) Middle layers are the most transferrable across tasks

(d) Final layers of BERT are more task specific.
- in pretraining — MLM task
If you after fine-tuning, if you restore the early layers to its original weight, it does not dramatic hurt the model performance.

6 Training

6.1 Pretraining

(a) Removing NSP does not hurt (sometimes improve performance

(b) Diverse masks during MLM pre-training within an epoch

(c) XLNet

6.2 Model Architecture Choices

(a) No. of heads not as important as no. of layers

(b) Large batch training (8k examples) improves.

(c) bsz = 32k, training time reduce with no perf-degree

(d) Normalize [CLS]

(e) Model-pruning by stacking - "warm start"

6.3 Fine-tuning BERT

(a) [SEP] attention on [SEP] basically tells BERT what to ignore. [SEP] attn increases on finetuning tasks.

(b) Propose using weighted repr of all layers instead of only the last layer.

(c) Adaptor modules - reduces the finetuning cost to a fraction of original

(1) How big should BERT be?

(1.1) Overparameterization

- [Pay less Attn with Lightweight & Dynamic Convolutions.] Feb 2019 - Good paper.

(a) Current models do not make good use of the parameters they already have.

Specialized heads do the heavy lifting - the rest can be pruned - 2019

What does BERT look at? Analysis of BERTs attention 201

Most layers could be reduced to a single head.

(b) Some heads are harmful for downstream tasks.

(c) Attn dropout could be the reason why some heads are redundant.

(1.2) BERT Compression.

Distillation

BigTeacher → SmallStudent

Read all the papers from mentioned here.

Quantization

- lower memory by lowering precision of weights.
- often requires compatible hardware.

what? [Decompose embedding matrix to small matrix
Progressive Model replacing

(8) Multi-lingual BERT

↕

(9) Discussion

(9.1) Limitations

(a) "the fact that a linguistic pattern is not there, doesn't mean that it is absent or the presence of one doesn't mean that it is used"
→ more complex probe might be able to recover more information. — BUT IT BECOMES LESS CLEAR WHETHER WE ARE TALKING ABOUT THE ORIGINAL MODEL.

(b) Ongoing debate about the merits of attention as a tool for interpreting deep learning models.
[WHAT? — On identifiability in Transformers, Attention is not explanation / Attention is not not explanation, Is Attention Interpretable?]

(9.2) Incentivize dataset development