ARTICLE COMMENTS NLP

PAPER TALK

Florian Laborde
AI Lab Intern - Telecom Paris
Etalab
DINUM
florian.laborde@data.gouv.fr

Pavel Soriano-Morales
Lab IA Datascientist
Etalab
DINUM
pavel.soriano@data.gouv.fr

Tam Kien Duong
Open data Team
Etalab
DINUM
tamkien.duong@data.gouv.fr

August 6, 2020

ABSTRACT

Qualitative analysis on article read. Transformers mainly. Written in 'oral english'. Some questions remain open. This is more of a personnal understanding than a formal summary.

1 Attention is all you need [1]

This is the groundbreak article in modern NLP. Attention models disrupt existing SOTA [2] papers based on RNN and LSTMs which have always been heavily flawed and contested model architectures. The attention layer which is presented is synonym with the resulting architecture of the 'Transformer'.

1.1 Articles to read to have full understanding

In all these new types of Neural architectures for NLP there still is of course some kind of tokenization [3]. It's a critical step in nlp and already a kind of embedding of the words. The subject is dense and relates to subwords algorithms or Neural Machine Translations. The mostly common used techniques are Byte Pair Encoding and WordPiece. BPE [4] can be seen as a very effective way of translating group of sounds/letters into numbers without really affecting the dimensionality or expressiveness of the initial sentences. The basic steps of the algorithm is a compression scheme. WordPiece [5] is more complicated to understand as the initial article is broad. But it is the same idea, just a little bit more sophisticated. The principle behind these algorithms is to regroup existing subwords together and allows for a limited vocabulary. (Like in German, many words are just a concatenation of other words). Didn't read them yet, but should to have a perfect understanding. WordPiece is used in the BERT model.

1.2 Questions ?

- What is different among the heads ? It is stressed on several times that the different heads from the 'multi-head' attention are outputting different views. What makes it so ? If architecture and Input/Output data are parallel it can't be. And I don't know which parameters influences this head differences ? Maybe it's just a random seed somewhere but can't find it in the code.

1.3 Interesting about

The annotated transformer is a Harvard website with a great code/ text walkthrough of the detailed implementation of a transformer network.

2 BERT: Bidirectional Encoder Representations from Transformers. [6]

El famoso Bert is the first big language model based on the transformer architecture. The real interesting questions are in the Appendix.

2.1 Summary

The main element in BERT is that it reads the text from left to right and right to left at the same time in the training process and allows for deeper bidirectional understanding of the sentences.

> BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

This leads to a unique Masked Language Model (MLM) pretraining which explainability for its effectiveness doesn't seem to be commented in the original paper. The overall pre-taining has a local aspect: MLM, and a global one: Next Sentence Prediction (NSP). The effect of both elements during pre-training is very well highlighted with the Ablation study from 5.1. Also, innovation is really around the pre-training/ fine-tuning protocol and the used datasets. It is pretrained on this MLM task (aka *Cloze* task) with objectives that aren't useful for NLP processing as such. The training tasks are really there to make the network learn 'language'. All of the applicative orientation is done only in fine-tuning. Although the general framework is really working on a pair of input of sentences: Q/A or Next sentence predictions etc.. A/B scenarios. The bigger the better works for BERT architectures whatever the scale task.

> we believe that this is the first work to demonstrate convincingly that scaling to extreme model sizes also leads to large improvements on very small-scale tasks

2.2 Questions ?

- Why does MLM work? It's strange that the network doesn't just remember or that random doesn't affect too much. Especially I have absolutely no understanding intuitively of the results of the last appendix table. The authors are very aware about this problem

  > bidirectional conditioning would allow each word to indirectly "see itself", and the model could trivially predict the target word in a multi-layered context

  and the induced mismatch of the MLM solution:

  > Although this allows us to obtain a bidirectional pre-trained model, a downside is that we are creating a mismatch between pre-training and fine-tuning, since the [MASK] token does not appear during fine-tuning.

and propose some interesting analysis and solution but fail to give a good overview of the Hows and Whys. The last sentence of the paper (Appendix)

> Interestingly, using only the RND strategy performs much worse than our strategy as well.

but they don't discuss it further.

- Didn't really get the Separators/Class problem or usefulness clearly.
- How do you deal with the input size in the network? Is it just masks? weird. (compared to images anyway)

2.3 Interesting about

About the Squad and Squad v2 datasets.SQuAD v1.1 does same thing as Piaf [7]

> Given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage.

and wonder if we could extend to a v2.0 that would also work when

> no short answer exists in the provided paragraph. Seems complicated after discussing it.

3 Sentence-BERT [8]

3.1 About BERT, transformers and the new paradigm in NLP relatively to embeddings

Why BERT changed the game? Well, we usually saw NLP as an embedding problem and then, do what you can with a correct representation. Now, with the power of the neural nets and BERT etc.. NLP is less embedding oriented. In fact, the embedding is an extremely difficult task because it is supposed to grasp the inner representation of language in a simple mathematical model. Which is most probably not possible as such. However, using the transformer architecture networks can just re-use existing language understanding. In the sense that you don't need to get down to the meaning of everything but just find what you need for your end task. A bit like a child would just repeat a whole sentence that gets him what he wants without having a deep meaning understanding of each word. Because of that, some NLP tasks produce amazingly good results for the human person.
But it's really like a database chatbot that's really clever, it does not mean that the network 'understands' what it is saying. (All this is not completely true and we see that in multi-head attention the words connections really show a deeper understanding of the sentence syntax but it's just to give an idea).
So, the BERT networks can produce whole new sentences and interact in a very clever way with the user. However, tasks such as similarity content do not require just to understand the input but also to have a very precise understanding of the whole language itself. For example, BERT can be good at paraphrasing and can generate one similar sentence, but its last layer embeddings are always related to the output of the task and can't be used to find an opposite sentence or anything different. So, we could guess that the BERT final layer embedding is not globally meaningful in language space. This is also due to the fact that as described and criticized in the BERT part, the model is always trained in a pair manner with the A/B paradigm, so embeddings for a single sentence don't make sense for themselves standalone.
Anyway, while BERT works quickly enough at inference for a multitude of complicated tasks, its architecture does not allow for repeated simple runs. As described in SBERT:

> Finding in a collection of n = 10 000 sentences the pair with the highest similarity requires with BERT n(n-)=2 = 49 995 000 inference computations. On a modern V100 GPU, this requires about 65 hours.

SBERT proposes an architecture to do change that. Also, what is great about SBERT is the training which focuses on having a globally meaningful sense. As we always thought, the locality of the embeddings in every training transformers or even Spacy-like GloVe etc.. embeddings is that you know that things are close together but not how far relatively to other things they are. In SBERT they use NLI databases of group of sentences that are either similar, opposite or neutral. This helps set points with larger distance among them and get globally meaningful embeddings.

3.2 Articles to read to have full understanding

You should read the triplet loss but the first article is really not clear about it, for this type of things as sad as it is, I think some towardsdatascience article would do a better job. Same thing for the Siamese networks which are presented in an antique article of LeCun and don't explain very much the problem. .

3.3 Questions ?

- Can we train it in French? probably not because the NLI dataset is really a hand-made thing. XNLI is a small dataset when using it a French language only and is designed to be used in the multilingual case. Although there is the thing with Wikipedia: why not?

4 SBERT - Multilingual [9]

The idea is just to use the info extracted from the previous NLI bases and then use multilingual sentence translations without worrying about any characteristic for the training because it just needs to translate the vector space. It's really something close to taking each point that's a word in one language and just translating it with a dictionary without changing space topology too much. Architecture is just a Knowledge distillation. It is a very good and practical idea to extend any understanding from a language on specifically built datasets, thus releasing the pressure for quality humanly specific-produced annotated data (like PIAF), using this from the English language and then just using translated texts. But, doing so all the real meaningful embedding always comes back to the initial NLI dataset. This loosen the constraint on the annotation and is a key idea in expanding NLP to different languages on specific tasks. One could say that the 'understanding' task is more complicated than the 'translation' task.
The main architecture is the block that does the embedding in English. This task is as described in SBERT. What we need is now to copy the same embedding for a different language. In this there are two elements:

- Copying the neural networks function (pure copy)
- Understanding new languages, translation task

As we need to teach a new model based on an initial one this gives the Teacher/Student architecture. The Loss is then computed based on the two requirements, we feed the translated sentences in every language and try to minimize output difference. One part of the loss is an MSE on the English language only to insure a correct embedding in the new model. The other part is an MSE on the embeddings between different languages sentences to insure a correct similarity between same sentences of different languages.

References

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser Illia Polosukhin: *Attention Is All You NeedAttention Is All You Need*. arXiv:1706.03762 [cs], . 2017. http://arxiv.org/abs/1706.037622020-07-01, arXiv: 1706.03762.

[2] Young, Tom, Devamanyu Hazarika, Soujanya Poria Erik Cambria: *Recent trends in deep learning based natural language processingRecent Trends in Deep Learning Based Natural Language Processing*. CoRR, abs/1708.02709, 2017. http://arxiv.org/abs/1708.02709.

[3] Webster, JonathanJ. Chunyu Kit: *Tokenization as the initial phase in NLPTokenization as the initial phase in NLP*. *Proceedings of the 14th conference on Computational linguistics* -, 4, 1106, Nantes, France, 1992. Association for Computational Linguistics. http://portal.acm.org/citation.cfm?doid=992424.9924342020-08-05.

[4] Sennrich, Rico, Barry Haddow Alexandra Birch: *Neural machine translation of rare words with subword unitsNeural Machine Translation of Rare Words with Subword Units*. CoRR, abs/1508.07909, 2015. http://arxiv.org/abs/1508.07909.

[5] Wu, Yonghui, Mike Schuster, Zhifeng Chen, QuocV. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes Jeffrey Dean: *Google's neural machine translation system: Bridging the gap between human and machine translationGoogle's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. CoRR, abs/1609.08144, 2016. http://arxiv.org/abs/1609.08144.

[6] Devlin, Jacob, MingWei Chang, Kenton Lee Kristina Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language UnderstandingBERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs], . 2019. http://arxiv.org/abs/1810.048052020-07-01, arXiv: 1810.04805.

[7] Keraron, Rachel, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, EdmundoPavel Soriano-Morales Jacopo Staiano: *Project PIAF: Building a Native French Question-Answering DatasetProject PIAF: Building a Native French Question-Answering Dataset*. *Proceedings of The 12th Language Resources and Evaluation Conference*, 5481–5490, Marseille, France, . 2020. European Language Resources Association, ISBN 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.6732020-07-27.

[8] Reimers, Nils Iryna Gurevych: *Sentence-BERT: Sentence Embeddings using Siamese BERT-NetworksSentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv:1908.10084 [cs], . 2019. http://arxiv.org/abs/1908.100842020-07-22, arXiv: 1908.10084.

[9] Reimers, Nils Iryna Gurevych: *Making Monolingual Sentence Embeddings Multilingual using Knowledge DistillationMaking Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. arXiv:2004.09813 [cs], . 2020. http://arxiv.org/abs/2004.098132020-07-21, arXiv: 2004.09813.