Research paper NLP

Master 2 Data Science

Transfer Fine-Tuning: A BERT Case Study

Tom Salembien

Professeurs :
Matthieu Labeau & Chloe Clavel

Contents

1 Highlight and Explanation

1.1 Bert Model

As an introduction, this article presents a new method to generate representations for semantic equivalence assessment tasks. To do so, Yuki Arase and Junichi Tsujii based their work on BERT's model [1] (Bidirectional Encoder Representations from Transformers). This encoder is a very powerful language representation model that marked a milestone in the field of automatic language processing, taking into account both left and right context of every word in the sentence to generate every word's embedding representation. BERT outperforms a large pannel of models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful.

1.2 Enhancement Idea

The guiding principle of the article is to improve BERT pre-training models for a specific class of NLP task which concerns the semantics equivalence of sentences. Yuki Arase and Junichi Tsujii, the authors of the article, demonstrate that sentential and phrasal paraphrase relations help sentence representation learning. They highlight a new BERT pre-trained model which is lighter in terms of parameters while improving usual base and large BERT pretrained model and semantic equivalence assessments tasks.Their method achieves performance gains while maintaining the model size. Furthermore, it exhibits improvement on downstream tasks with limited amounts of training datasets for fine-tuning, which is a property crucial for transfer learning.

1.3 Paraphrases Analyse

The first research track is about the analyse of paraphrases in term of semantics. Yuki Arase and Junichi Tsujii have written an other article highlighting a method to obtain phrasal paraphrases based on phrase alignment on parse forests [2]. The advantages of this method is that it identifies syntactic paraphrases under linguistically motivated grammar and it allows phrases to non-compositionally align to handle paraphrases with non-homographic phrase correspondences. We finally get a dataset with parse trees and their phrase alignments.

1.4 Transfer Fine-Tuning with Paraphrasal Relation Injection

The method uses two types of paraphrase classification, phrasal and sentential which complements each other on sentence representation learning. The objective of phrasal paraphrase classification is to give an explicit supervision of semantic relations among phrases and a structures in sentences through representation learning. The objective of sentential paraphrase classification is to discriminate the paraphrase sentence pair from random. The initial work is to we designed a three-way classification task: discrimination of paraphrase, random, and in-paraphrase so that the network learns to discriminate sentences that contain a large number of identical or related words. Then these pairs are fed into the following algorithms.

Algorithm 4.1 Paraphrasal Relation Injection

Input: Paraphrase sentence pairs $P = \{\langle s, t\rangle\}$, a pre-trained BERT model

Obtain a set of phrase alignments A as pairs of spans for each $\langle s, t\rangle \in P$
WordPiece tokenization of P
Accommodate phrase spans in A to BERT's token indexing: $A = \{\langle (j, k), (m, n)\rangle\}$
repeat
  for all mini-batch $b_t \in \{\langle P_i, A_i\rangle\}$ **do**
    Encode b_t by the BERT model
    Compute loss: L(Θ)
      For phrasal paraphrase task: $L_p(\Theta)$
      For sentential paraphrase task: $L_s(\Theta)$
      $L(\Theta) = L_p(\Theta) + L_s(\Theta)$
    Compute gradient: $\nabla(\Theta)$
    Update the model parameters
until convergence

The first three steps constists in tokenisation[3] and extract phrasal relations with their alignment method. Then for the phrasal paraphrase classification, the BERT model encodes the tokenized pairs to get the hidden states. Then, they reduce the dimension and extract the most representative features with a max-pooling layer, we get h_s, h_t. Finally, they extract the relations $([h_s, h_t], h_s \odot h_t, |h_s - h_t|)$ and input them in a classifier, a simple Fully-Connected layer and softmax.

For the sentential paraphrase classification, it is the same process until BERT, then the first hidden stateh_1 (for [CLS]) is fed into a Fully-Connected layer and softmax.

1.5 Main hypothesis verified empirically on GLUE

There are five main hypothesis that will be verified with GLUE (General Language Understanding Evaluation [4]) standard benchmark.

H1	Our method contributes to semantic equivalence assessment tasks.
H2	Our method achieves improvement on downstream tasks that only have small amounts of training datasets for fine-tuning.
H3	Our method moderately improves tasks if they are relevant to semantic equivalence assessment.
H4	Our training does not transfer to distant downstream tasks that are independent to semantic equivalence assessment.
H5	Phrasal and sentential paraphrase classification complementarily benefits sentence representation learning.

The article shows that Semantic Equivalence Assessment Tasks (MRPC, STS-B, QQP) verifies **H1**. Several test has been done regarding the size of the training sets (table 4 of the article), they show that their fine-tuned model performs better than BERT-Base when the number of data for training is limited on downstream tasks, which is a property desirable for transfer learning and which supports hypothesis **H2**. The natural language inference (NLI) Tasks (MNLI-m/mm, RTE, QNLI) verifies **H3**. Single-Sentence Tasks (SST, CoLA) verifies **H4**. Finally for the last hypothesis, an ablation study highlights independent effects of sentential and phrasal paraphrase classification. With the same initialization and training, the table 5 of the article shows that phrasal and sentential paraphrase classification are complementary, **H5**.

1.6 Pros and cons of the proposed method

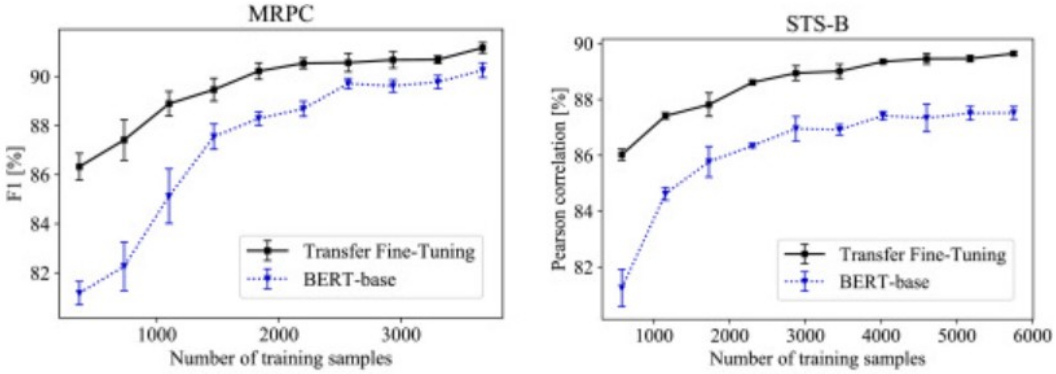
I will start with the cons, the main criteria that we can make on this model is that it improves Bert model only in semantic similarities of the sentences, it is not transferable to down stream tasks independent from semantic analysis. The performance increases but the generality of the model on several tasks decreases. An other issues is that the performance of the model also depends on the domain of the corpora, in fact, they showed that the domain mismatch between our paraphrase corpora and QQP corpus may affect the performance.

But on the other hand, this article presents many advantages cited in the hypothesis. For a

smaller size, the model globally outperforms BERT-large on semantic equivalence tasks. This method also allows the generation of representations for semantic equivalence assessment tasks. Finally, this method shows an important propriety for transfer learning, with limited size of datasets for training for fine-tuning the models seems to be efficient.

2 Evolution that followed

The paper let many questions and tests to do. Indeed, even the authors of the article were saying that in the future they will have to test different sizes of BERT models and improve the alignement method quality. That is why in 2021, YukiArase and Junichi Tsujii wrote "Transfer fine-tuning of BERT with phrasal paraphrases" [5]. In this new article, they studied the convergence of the model depending on the size of the corpus doing several test on the GLUE benchmark and the PAWS dataset, comparing the new model with the precedent one and also BERT-base and large. As a consequence, their analyse is more robust as they used several databases (GLUE and PAWS) to test and compare the model. Thus, it fulfil the issue with the influence that might have the subject of the corpus on the performance of the model.



(a) MRPC (semantic equivalence)

(b) STS-B (semantic equivalence)

They also highlighted that simple features outperform elaborate ones in phrasal paraphrase classification. Indeed, the 2019 model used Conneau et al. (2017)[6], in which the features extracted were pretty complex to perform well. Now in the Arase and Tsujii (2021) model they use as simple a feature generation assuming that simple features are really efficient transmitting learning signals to the underlying BERT model.

	Pooling	Matching
(Arase and Tsujii, 2019)	Max-pooling	$(\mathbf{h}_s, \mathbf{h}_t, \mathbf{h}_s * \mathbf{h}_t, \lvert\mathbf{h}_s - \mathbf{h}_t\rvert)$
Our method	Mean-pooling	$(\mathbf{h}_s, \mathbf{h}_t)$

That is why mean-pooling is used to generate representations of source and target phrases $(\mathbf{h}_i)_i$ instead of the several element wise operation.
With an alignement <(j,k),(m,n)> and $2 \leq j < k < m < n \leq N-1$

$$\mathbf{h}_s = (\mathbf{h}_j, ..., \mathbf{h}_k)$$

$$\mathbf{h}_t = (\mathbf{h}_m), ..., \mathbf{h}_n)$$

Fine-tuning influences a lot BERT's performance in downstream tasks. Liu et al. (2019)[7] proposed a method related and complementary to Arase and Tsujii (2019) model with fine-tuning in a multitask learning setting. That is why the 2021 model discriminates phrasal and sentential paraphrases simultaneously by multitask learning and improves the performances of the model on several tasks.

Finally, Arase and Tsujii (2019) article also opened several approaches based on the injection of external information into neural networks to generate representations dedicated to a specific purpose. Here are few ones, SentiBERT (Yin et al., 2020)[8], sentiment information injection into BERT, for sentiment analysis task. ERNIE (Zhang et al., 2019b) [9], entity knowledge injection for sentence representation.

3 Personal take on the interest and possible improvement

This article is really interesting, indeed it uses fine-tuning to increase the performance of BERT model on a specific task while keeping the same size. Also the improvement on limited training dataset is also a real improvement concerning transfer learning. Hovewer, in my opinion few enhancement might have been done.

The self-attention mechanism allows to extract keys context vector that focuses attention on input and it also allows inputs to interact with each other. A possible upgrade to this is that the attention coefficient are 1D vectors, they are summarizing the information of the sentence. Thus,

a one dimensional vector might be a bit restrictive. The self-attention mechanismis based on the following equations,

$$u_t = tanh(Wh_t) \quad \alpha_t = \frac{exp(u_t^T u}{\sum_{t'=1}^{T} exp(u_{t'}^T u)} \quad s = \sum_{t=1}^{T} \alpha_t h_t$$

The idea of the article [10] is to enhance the self attention mechanism pruning weight connections using a 2D matrix instead of a vector for the context. They also propose to add a penalization term as the embedding matrix M suffers from redundancy to encourage the diversity of summation weight vectors across the attention mechanism. With a 2D matrix structure and a penalization term the model will have a greater capacity to extract the latent information from the input sentences. Now in the enhanced model, we introduce H = $(h_1, h_2, ..., h_n)$

$$U = tanh(W_{s1} H^T) \quad A = softmax(W_{s2} U) \quad S = AH$$

4 Implementations and tests

There is no implementation available, but after some researches here is the link the Arase github project for transfer fine-tuning. She only exposes the weights matrix of the model. Thus, I tried to implement BERT model and test it on GLUE datasets (specifically on MRPC). But for now I had few issues with my computer load the libraries, I also tried on colab but it takes a lot of time to load the 'transferFT-bert-base-uncased.pkl' pkl file containing the weights of there pretraining. Here is the link to my implementation from hugging face : tom-implementation

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[2] Y. Arase and J. Tsujii, "Monolingual phrase alignment on parse forests," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1–11. [Online]. Available: https://aclanthology.org/D17-1001

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. [Online]. Available: https://arxiv.org/abs/1609.08144

[4] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJ4km2R5t7

[5] Y. Arase and J. Tsujii, "Transfer fine-tuning of bert with phrasal paraphrases," *Computer Speech Language*, vol. 66, p. 101164, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300978

[6] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 670–680. [Online]. Available: https://aclanthology.org/D17-1070

[7] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4487–4496. [Online]. Available: https://aclanthology.org/P19-1441

[8] D. Yin, T. Meng, and K.-W. Chang, "SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3695–3706. [Online]. Available: https://aclanthology.org/2020.acl-main.341

[9] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1441–1451. [Online]. Available: https://aclanthology.org/P19-1139

[10] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *CoRR*, vol. abs/1703.03130, 2017. [Online]. Available: http://arxiv.org/abs/1703.03130