NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

Saint Petersburg School of Physics, Mathematics, and Computer Science

Department of Computer Science

PROJECT PROPOSAL:

Preprocessing Sequential Data for Machine Learning Facilitation

Maxim K. Surkov, group BPM171

Linguistic Advisor:

Department of Foreign Languages

Research Advisors:

Ivan P. Yamshchikov,

Associate Professor,

Department of Informatics

Ph.D. in Physico-Mathematical Sciences

St. Petersburg

2021

Modern state-of-the-art natural language processing systems use deep neural networks (e.g., BERT, GPT-3) that require many resources during both training and inference. Several optimization techniques have been developed for the last ten years. One of them is curriculum learning, which consists of two parts, namely data complexity evaluation and sampling. The purpose of our work is to research existing metrics of text complexity, develop new effective ones to reduce the learning time of deep neural networks on pre-training and classification tasks, and make a comparative analysis on them. In this work, we consider already existing metrics, such as length and TF-IDF. Also, we found two expressive metrics (Excess Entropy, TSE) and adapted them for text complexity evaluation. Finally, we compare the above metrics on pretraining and classification tasks using the BERT deep neural network and research the influence of presented metrics on training time. It is expected that there exists an effective training curriculum based on given metrics. This work is assumed to introduce a new text complexity evaluation method, which will allow to speed up BERT convergence time.

Keywords: natural language processing, curriculum learning, information theory

Introduction

Background.

There is an enormous number of fields where natural language processing is applied in the modern world, such as artificial voice assistants, text filtering in messengers, machine translation, etc. The tasks that arise within these systems must be solved efficiently and accurately.

Fortunately, for the last three years, the attention mechanism has been developed (Vaswani et al., 2017). Based on this technology, machine learning engineers created BERT and GPT-3 models to solve the given tasks with very high accuracy (Devlin et al., 2019; Brown et al., 2020). But at the same time, these models require a significant amount of time to train. It is worth adding that the training process usually takes place using expensive graphics processors. For instance, BERT pre-training takes from a few days to several months. Interestingly, the world record for BERT pre-training of 47 minutes was set by the Nvidia company, where scientists used about 1500 video cards. Simultaneously, the BERT model fine-tuning requires several days, even on elementary tasks such as spam or hate speech classification (Narasimhan et al., 2020).

Recently, several methods have been developed to facilitate the BERT model. One such technique is curriculum learning which is actively used in various machine learning fields with great success, but in natural language processing, there are a limited number of remarkable results (Platanios et al., 2019; Xu et al., 2020). In this set of works, the influence of curriculum learning on machine translation and classification tasks has been already discovered. But authors of such papers pay great attention to sampling algorithms while the complexity metrics remain unexplored. The text length, which is a metric of the current best

curriculums for downstream tasks, is not expressive enough for text difficulty estimation. As a result, we have a large field for research on this significant issue.

Problem Statement.

Our work's primary purpose is to speed up the BERT model's training process at the expense of applying effective text complexity estimation metrics within the framework of curriculum learning on pre-training and classification tasks. It is necessary to solve the following problems to achieve this goal.

Firstly, we should suggest alternative text complexity metrics that will be potentially expressive and effective in terms of the BERT convergence speed. Moreover, we have to implement a practical algorithm for metrics calculation on large datasets of several million samples.

Secondly, a comparative analysis should be carried out between the proposed metrics and the existing ones. The comparison's main characteristic is the relative position of the model accuracy graphs as a function of the number of training steps. The metric that shows fewer steps to achieve a particular accuracy is more efficient with the fixed model, dataset, and sampling algorithm.

The final problem is to study the impact of the found metrics on the BERT training time on pre-training and classification tasks. The comparison algorithm is the same as in the previous problem, except that we fix concrete metrics and use several different sampling algorithms to highlight the text complexity estimation method's influence on the convergence speed.

Professional Significance.

An efficient curriculum with expressive enough text complexity estimation metrics allows us to train deep neural networks much faster. As a consequence, much cheaper in terms of using graphical processors, which are very expensive. Moreover, it might help scientists in the natural language processing field to conduct more experiments and test more hypotheses in less time.

The method of adapting information theory metrics to language processing presented in this work should extend the existing tools for assessing text documents' complexity.

Delimitations of the Study.

As we have mentioned before, the main models for natural language processing tasks are BERT and GPT-3, but we work only with the first one. There are several reasons for this. Firstly, the GPT-3 is extremely large in terms of memory usage, and we are not able to work with such a model comfortably. Secondly, the BERT model has a much higher quality prediction ability than the smaller models (for example, multilayer LSTM).

Structure of the Paper.

The paper is structured in the following way:

- section "Literature Review" describes the existing approaches of curriculum learning applications and text complexity estimation metrics
- section "Methods" introduces our methods inspired by information theory
- section "Results Anticipated" highlights the main conclusions obtained in the course of our study

Literature Review

Origins.

The crucial issue of training time in natural language processing has been examined intensively for several years. Sheng Shen et al. (2019) suggested a new method of compressing the BERT model size using quantization techniques. Limiting the number of possible values of language model weights, they reduced BERT's size several times without significant accuracy loss. About a year later, Hassan Sajjad et al. (2020) conducted a series of experiments related to dropping different subsets of layers from the BERT and showed that it is possible to remove up to 40% of the model parameters in such a way that the BERT does not stop showing satisfactory results on the main tasks of natural language processing. Overall, by modifying the neural network architecture, the approaches mentioned above allow us to get compressed versions of the large model that can converge much faster.

These kinds of techniques invade the network's inner structure and require additional research for specific models. Simultaneously, there is another way to speed up the models' convergence called curriculum learning. Unfortunately, it is not known precisely when the idea of applying this technology in machine learning was first proposed, but it can be argued that the logical beginning is the paper of Bengio et al. (2009). It was shown that curriculum learning leads to improving machine learning model quality. The authors set up several experiments, one of which was about the classification of geometric shapes. They found out that if you show the simple examples to the model (squares, circles, and equilateral triangles) before the primary training cycle, the final model will be better in terms of quality. This simple example shows that curriculum learning has a great potential for improving existing solutions in machine learning.

Applications in Related Machine Learning Fields.

Curriculum learning has been actively applied in different machine learning areas for the last several years. For instance, Hacohen and Weinshall (2019) came up with the idea of a curriculum for image classification problems in computer vision. They suggested a model-based metric for the images, which can be calculated in the following way. Firstly, it is necessary to take an independent convolutional neural network, pre-trained on the ImageNet dataset. The second step is defining image complexity as the model's confidence score in its prediction in the given image. Finally, scientists used a ladder-like sampler paired with the developed complexity metric. As a result, the suggested approach increased both learning speed and final model quality.

Studies on the impact of curriculum learning were conducted earlier. Mermer et al. (2017) developed a curriculum in which sample complexity was calculated using ensemble and clustering methods. The authors showed a huge capability of curriculum learning in classical Deep Learning.

Applications in NLP.

In natural language processing, there exists a limited number of significant results. Perhaps this is because natural languages consist of words and sentences that do not have a clear structure that can be properly formalized. Moreover, unfortunately, science cannot clearly describe the processes inside modern language models by which we have high-quality solutions.

Platanios et al. (2019) investigated curriculum learning's influence on the machine translation task's language models convergence speed. The authors considered two metrics of text complexity, namely the length and minus logarithm of likelihood probability. They showed that the latter does not provide any profit in terms of model training speed than the

former. But scientists did not conduct a study on finding a better metric for text complexity estimation. Moreover, they managed to achieve a good result only on the machine translation while nothing said about other types of tasks.

Another approach to applying curriculum learning in natural language processing was suggested by Xu et al. (2020). Their method operates with such a concept as model-based text complexity estimation. It means that the difficulty of samples changes during the training process according to how well the model has trained at the current time. In more detail, the scientists use a complicated algorithm that consists of several steps. In cycle, the following process is repeated. First of all, the training dataset is split into several parts. After that, a free-standing, independent BERT is trained initialized with the current global model's weights for each piece. Using these prepared models, we can estimate the text samples' complexity as a sum of confidence scores of BERTs. Finally, the authors re-sort the whole dataset according to the calculated complexities and repeat the process until convergence. Their method allows us to improve final model quality on classification tasks significantly. Still, the described approach is much more resource-intensive and does not provide any mechanism for decreasing language model training time. As a result, we do not have the opportunity to compare their technique with ours.

Existing Text Complexity Metrics.

On the first side, it seems that it is possible to develop an infinite number of different metrics that estimate text data complexity. Still, it turns out that it is not obvious at all how to correctly formulate the concept of complexity for sequences, especially for the text. M. Zakaria Kurdi et al. (2020), in their paper, considered several dozen ways to estimate text complexity, which can be divided into the following groups: phonological, morphological, lexical, syntactic, discursive, and psychological metrics. This work has been done to deal

with the problem of building “a classifier that can identify text complexity within the context of teaching reading to English as a Second Language (ESL) learners” (Kurdi, 2020, p. 1). The scientist operated with understandable terms for humans, but not for deep neural networks such as nouns, adjectives, verbs, etc. As a result, we can assume various useful metrics, distinguishing the most complex samples from the easiest ones. But, Frans van der Sluis and Egon L. van den Broek (2010) showed that complexity estimation approaches mentioned above poorly correlate with real text complexity. They examined two versions of Wikipedia: the usual and simple one. The main difference between them lies in the length and content of the articles these datasets consist of. The authors compared readability metrics based on a syntactic characteristics calculation (for example, words per sentence) with entropy and semantic complexity, calculated using the WordNet database. The scientists found out that while length and readability can not distinguish between complex and easy Wikipedia articles, entropy and semantic complexity have strong “differentiation power between both Wikipedias” (van der Sluis et al., 2010, p. 4). Consequently, it makes no sense to consider classical metrics for text complexity estimation based on the main syntactic characteristics.

Tom Kocmi et al. (2017) and Xuan Zhang et al. (2018) independently showed that the length and word frequency rank-based metrics are expressive enough to decrease language models’ training time on machine translation tasks. As a consequence, we have to consider them in further investigation.

The text data complexity estimation issue can be explored from an information theory point of view. Fortunately, there exists a paper by Nihat Ay et al. (2006), where authors developed “A Unifying Framework for Complexity Measures of Finite Systems” (2006, p. 1). They suggested four complicated and expressive metrics, two of which can be adapted to our task, namely Excess Entropy and TSE.

Methods

To achieve the global goal, we set up a series of problems to be solved. The first one was to suggest alternative metrics for text complexity estimation, which are potentially expressive enough to help us speed up the BERT model's convergence, and implement an efficient algorithm to calculate the found metrics on large datasets. As we mentioned before, several metrics were already used in related works, such as length and word frequency-based metrics. We do not use classical metrics for text complexity estimation based on syntactic characteristics because such approaches do not have a solid ability to distinguish complex examples from simple ones (van der Sluis et al., 2010). Besides, we consider one more method that is popular in information retrieval, namely TF-IDF. Several very expressive metrics are taken from information theory: excess entropy (EE), TSE (Ay et al., 2006). We develop a methodology of transforming the text into a jointly distributed random value necessary for calculating information theory based metrics (EE and TSE). We tokenize text using BPE (Byte Pair Encoding) model as is usually done when using the BERT model. Each number in the sequence obtained in the previous step generates a binary random value that depends on its weight and position in the sequence. Next, we concatenate generated random values into one vector, which is the jointly distributed random variable we need. After that, we use well-known mathematical equations and make assumptions about the text's far-placed parts' independence to simplify the metric calculation. Finally, to make it possible to compute suggested metrics on a large dataset, we have to collect different statistics from them (for instance, the number of samples with a particular length) efficiently. To face this issue, we divide the dataset into several parts, process them in parallel on several processors, and merge accumulated data into one place.

The second goal was to make a comparative analysis of the gained metrics among themselves. First of all, we should set a configuration of experiments accurately. We use

sentiment140 and HND datasets which contain training samples for complicated tasks on which the BERT model would train long enough for the application of acceleration methods to be reasonable. We fix the dataset, model, and sampling algorithm for a clean comparison of two different metrics. After that, we train the BERT model using two given curriculums based on the considered metrics. To understand which approach is better, we analyze the model's convergence graphs, fixing a sufficiently large accuracy value and comparing the number of steps required to reach this threshold for each of the metrics.

The final task is to explore curriculum learning's influence with suggested metrics on the BERT model convergence speed. To solve this problem, we fix particular metrics and compare the curriculum-based approach with the baseline in the same manner as in the previous task solution.

Results Anticipated

We have chosen several expressive metrics which have a potentially strong ability to distinguish complex data samples from simple ones. Moreover, we have developed and implemented an efficient algorithm that calculates suggested metrics on a large amount of data. Our statistics calculation approach can be scaled into extremely large datasets with the condition of having a sufficient number of processors. Also, our metrics calculation algorithm allows us to estimate text complexity in linear time, which is crucial because the first version of the difficulty evaluation method required obscenely large resources.

A comparative analysis of the metrics revealed a series of critical observations. Firstly, if we use a trivial sampling algorithm that sequentially retrieves examples in the sort order, it turns out that the text complexity estimation metric does not affect anything. Secondly, in the case of using a sampling algorithm by Platanios et al. (2019), it can be seen that the exist statistically significant differences between the metrics. TF-IDF, Excess Entropy

and TSE can achieve a reasonable accuracy of 84.5% in about 20k training steps (EE and TSE based curriculums behave more stably). At the same time, the length-based method does it in approximately 35k iterations. Thereby, we can conclude that more expressive estimation techniques give a training acceleration of 43% on classification tasks. But it depends heavily on the sampling algorithm. For instance, if we sample examples using a sliding distribution window, the length's effect will speed up the training more than TF-IDF does. However, TSE and EE will remain in the lead, showing a slightly less increase in convergence speed of 32% compared to the length. In total, experiments on using such metrics as length, TSE, EE, and TF-IDF were conducted on two datasets (HND and sentiment140) that relate to the classification problem. Also, similar research has been done on pre-training tasks using the Wikipedia dataset, but, unfortunately, we noticed a bizarre behavior of the model on this task type. Finally, we still have to explore the influence of the remaining metric (word frequency-based approach) on the BERT convergence speed both on pre-training and classification tasks.

Conclusion

In the fast-paced world of high technologies, natural language processing is crucial for the vast majority of applications. The problems that arise in this area are solved with high-demand deep neural networks, which require a large amount of time to be trained. As a result, the training acceleration issue is crucial. One of the promising methods is curriculum learning that has already proven itself well in related machine learning fields. However, this technique's influence on huge language models' convergence speed is not well understood yet. Curriculum learning consists of two parts, namely data complexity estimation and sampling. Unlike, for instance, images on real-valued vectors, it is a complicated task to distinguish complex texts from simple ones, mainly since word sequences do not have a clear

structure. As a result, we have only a few basic metrics that are actively used in various curriculum learning applications.

We found several expressive metrics in this work, which gave us a statistically significant reduction in language model training time in some configurations. Moreover, we provided an efficient algorithm for suggested metrics calculation in a reasonable amount of time. Also, we have conducted a comparative analysis of text complexity estimation approaches and showed that the effect of a particular method depends on the selection of the sampling algorithm. However, some of the metrics (EE and TSE) perform best regardless of the experiment configuration. We plan to extend a described analysis into a wider set of metrics and NLP tasks.

References

Ay, N., Olbrich, E., Bertschinger, N., & Jost, J. (2006, August). A unifying framework for complexity measures of finite systems. In *Proceedings of ECCS* (Vol. 6).

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41-48).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Hacohen, G., & Weinshall, D. (2019, May). On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning* (pp. 2535-2544). PMLR.

Kocmi, T., & Bojar, O. (2017). Curriculum learning and minibatch bucketing in neural machine translation. *arXiv preprint arXiv:1707.09533*.

Kurdi, M. Z. (2020). Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL. *arXiv preprint arXiv:2001.01863*.

Mermer, M. N., & Amasyali, M. F. (2017). Scalable Curriculum Learning for Artificial Neural Networks. *IPSI BGD TRANSACTIONS ON INTERNET RESEARCH*, *13*(2).

Narasimhan, S., Narasimhan, V. A. P. B. S., Karch, G., Rao, R., Huang, J., Zhang, Y., Ginsburg, B., Chitale, P., Sreenivas, S., Mandava, S., Ginsburg, B., Forster, C., Mani, R., & Kersten, K. (2020, October 13). *NVIDIA Clocks World's Fastest BERT Training Time and Largest Transformer Based Model, Paving Path For Advanced Conversational AI*. NVIDIA Developer Blog. https://developer.nvidia.com/blog/training-bert-with-gpus/

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., & Mitchell, T. M. (2019). Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.

Sajjad, H., Dalvi, F., Durrani, N., & Nakov, P. (2020). Poor Man's BERT: Smaller and Faster Transformer Models. *arXiv preprint arXiv:2004.03844*.

Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., & Keutzer, K. (2020). Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8815–8821. https://doi.org/10.1609/aaai.v34i05.6409

van der Sluis, F., & van den Broek, E. L. (2010, August). Using complexity measures in information retrieval. In *Proceedings of the third symposium on information interaction in context* (pp. 383-388).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Xu, B., Zhang, L., Mao, Z., Wang, Q., Xie, H., & Zhang, Y. (2020). Curriculum Learning for Natural Language Understanding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6095–6104. https://doi.org/10.18653/v1/2020.acl-main.542

Zhang, X., Kumar, G., Khayrallah, H., Murray, K., Gwinnup, J., Martindale, M. J., ... & Carpuat, M. (2018). An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.