CANINE : Pre-training an Efficient Tokenization-Free Encoder for Language Representation:
Project proposal

Chloé Sekkat
ENSAE Paris & ENS Paris-Saclay
chloe.sekkat@ensae.fr

Jocelyn Beaumanoir
ENSAE Paris & ESSEC
jocelyn.beaumanoir@ensae.fr

Abstract

This project is part of the Machine Learning for Natural Language Processing course at ENSAE Paris. We will work on a tokenization-free encoder for language representation called CANINE. It is the first character-level pre-trained deep encoder that is completely tokenization-free and vocabulary-free. We will study the performance of CANINE on several NLP tasks and compare it to other transformer-based models.

1. Problem Definition

Transformers have revolutionized the field of NLP, with remarkable generalization through transfer learning. However, most of these systems require an explicit tokenization step, which poses many limitations. First, subword tokenizers struggle with noisy text [11], and their method of splitting can also be unsuitable for certain morphologies of languages. Moreover, tokenizers use a fixed vocabulary during pre-training, and this limits their generalization capacity for downstream tasks, especially in specialized domains [2]. These limitations question the need for tokenization. Character-level models, that directly operate on characters instead of tokens, have emerged a while ago, but their performance was poor compared to their word-level counterparts. CANINE [5] is the first pre-trained tokenization-free encoder that outperforms BERT on some specific tasks. It operates directly on character sequences, and uses a downsamlping strategy to reduce the input sequence length, followed by a deep transformer encoder. In this project, we want to study the performance of CANINE in comparison with BERT on several NLP tasks.

2. Experiments

2.1. Models

First, we will make sure to understand how character-level language models work and differ from token-based language model, focusing on CANINE. Then, the latter will be compared against BERT [6] and mBERT (multilingual version) which are very powerful models often reaching state-of-the-art on many downstream tasks. Additionally, we might compare it to models encompassing both word-level and character-level information such as CharacterBERT [3] and ByteT5 [12]. Pre-trained versions are available thanks to Hugging Face.

2.2. Datasets

- Question Answering (QA): experiments will be done using reading comprehension datasets such as SQuAD [9], FQuAD [8] (unilingual) and XQuAD [1] (multilingual) in order to test the abilities and limits of CANINE, especially when training on datasets smaller than TYDI QA [4] (we will take only a subset of the datasets we'll use in order to restrict the computational resources needed).
- Sentiment Analysis: a common dataset for such task is SST-2 [10], which allows for a fine-grained classification between 5 levels of sentiment varying from very negative to very positive, and also for a binary classification between positive and negative. We will also test the strengths of CANINE in a less structured text with the sentiment140 [7] dataset, which is composed by 1.6 million tweets.

3. Evaluation

We will focus on 2 downstream tasks: QA (F1-score) and Sentiment Analysis (Accuracy). We will use the provided pre-trained checkpoints, both with autoregressive character loss and subword loss, in order to finetune them for each given task. Note that we would like to focus on exploring the datasets (exploratory data analysis in order to check whether the principles highlighted in the lectures also apply to ours) and the question answering task (especially since we want to finetune it on several datasets and that it is computationally heavy).

References

[1] M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

[2] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. *arXiv preprint arXiv:2010.10392*, 2020.

[3] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters, 2020.

[4] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 2020.

[5] J. H. Clark, D. Garrette, I. Turc, and J. Wieting. CANINE: Pre-training an efficient tokenization-free encoder for language representation, 2021.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[8] d. Martin, V. Maxime, B. Wacim, and B. Tom. FQuAD: French Question Answering Dataset. *arXiv e-prints*, page arXiv:2002.06071, Feb 2020.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250, 2016.

[10] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[11] L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, and C. Xiong. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*, 2020.

[12] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models, 2021.