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

Yujin Cho
ENS Paris-Saclay
yujin.cho@ens-paris-saclay.fr

Chloé Sekkat
ENSAE Paris
chloe.sekkat@ensae.fr

Gabriel Baker
Telecom Paris
gabriel.baker@telecom-paris.fr

Loubna Ben Allal
ENS Paris-Saclay
loubna.ben_allal@ens-paris-saclay.fr

Abstract

This project is part of the Algorithms for Speech and Natural Language Processing course of the MVA master. We will work on a tokenization-free encoder for language representation called CANINE. It is the first character-level pretrained 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 BERT.

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 [12], 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 [15]. Pre-trained versions are available thanks to Hugging Face.

2.2. Datasets

- Question Answering (QA): we will use the TYDI QA [4] (multilingual) dataset to reproduce the paper's results. Additional 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 on low resources languages and/or morphologically rich ones. Finally, one might want to evaluate CANINE on noisy QA datasets to see how it would perform in non-ideal settings.
- Name Entity Recognition (NER): we will use CONLL-2002 [13] and CONLL-2003 [14]. The main task of theses datasets consists on language-independent named entity recognition. It focuses on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. We will compare and test the robustness of CANINE with BERT model.
- Sentiment Analysis: a common dataset for such task is SST-2 [11], 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 twitters.
- Medical Natural Language Inference (NLI): we expect CANINE pre-trained model to have better generalization

than tokenization-based models to specific domains, we want to test this hypothesis in the medical domain. We will compare CANINE to BERT on the MEDNLI dataset for Natural Language Inference [10]. We will also compare it to CharacterBERT. In this task, we want to classify sentence pairs into three categories: Contradiction, Entailment, and Neutral. We will also evaluate the robustness of the model to noisy text.

3. Evaluation

We will focus on 4 downstream tasks: QA (F1-score), NER (Precision, Recall, F1-score), Sentiment Analysis (Accuracy), and Medical NLI (F1-score). 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.

4. Task sharing and Plan of work

Each student will work on different evaluation tasks. C.Sekkat will work on Question Answering, G.Baker will work on Sentiment Analysis, Y.Cho will perform evaluation for Named Entity Recognition, and L.Ben Allal will work on Medical Natural Language Inference.

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] A. Romanov and C. Shivade. Lessons from natural language inference in the clinical domain.

[11] 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.

[12] 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.

[13] E. F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[14] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[15] 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.