data2vec : A General Framework for Self-Supervised Learning in Speech, Vision and Language: Project proposal

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

Mathilde Kaploun
ENSAE Paris
mathilde.kaploun@ensae.fr

Abstract

This project is part of the Deep Learning course at ENSAE Paris. We will work on the first multi-modal self-supervised algorithm: data2vec [2]. It has been developed by FAIR quite recently and has shown promising results in speech, NLP and computer vision. Our goal will be to study its performance on several tasks in NLP and Computer Vision and compare it to the best single-purpose algorithms in each domain.

1. Problem Definition

One challenge of self-supervised learning is that all developed methods are unimodal. This means that each scheme is task-specific, e.g. Masked Language Model for NLP, learning representations invariant to data augmentation in Computer Vision. Consequently, attached to each task, are biased in their own way. Baevski et al. [2] are motivated by the core idea that humans use similar processes to interacted and understand the visual world, language and sound. Is it possible to design an algorithm inspired by such processes ? An algorithm that can learn a contextualized latent representations instead of modality-specific representations. Concretely, the learning objective in DATA2VEC is the same across all modalities: masked learning to produce latent target *continuous and contextualized representations*, using a teacher-student architecture training scheme.

2. Experiments

2.1. Models

First, we will make sure to understand how DATA2VEC work and differ from other self-supervised models in NLP and Computer Vision. Then, the DATA2VEC will be compared against SOTA models such as BERT [5] or RoBERTa [7] in NLP, and ViT [6] in Computer Vision on a various set of downstream tasks. Results provided by the paper suggest that DATA2VEC [2] overperforms such models. It would be nice (if pre-trained models are available) is to compare DATA2VEC with DINO [3] since both of them use the teacher-student architecture but differ in the prediction task.

Please note that currently pretrained models for Computer Vision are **not** available [1]. Depending on when FAIR will make one available, we might **not** be able to evaluate DATA2VEC on downstream Computer Vision tasks. Moreover, note that currently fine-runed pre-trained models on NLP tasks are **not** available, which means that we will fine-tune ourselves the base model. This will allow us to get a better understanding on how to fine-tune such models and will make us dive into the source code.

2.2. Datasets

- Computer Vision: as we will try to first reproduce the paper's results, we will use the well-known benchmark ImageNet [4], on the downstream tasks of image classification. Other tasks might be considered such as image captioning using the COCO (Common Objects in Context) dataset [2].
- Natural Language Processing: in order to reproduce the paper's results, we shall fine-tune the model on the GLUe benchmark. Note that not all tasks might be considered e.g. we might consider only Question Answering (SQUAD [8]), Natural Language Inference (MNLI [10]) and Sentiment Analysis (SST-2 [9]). One interesting thing would be to evaluate DATA2VEC performances on multi-lingual Question Answering for instance (using XQUAD [1]).

3. Evaluation

For image classification we will focus on top-1 validation accuracy on ImageNet-1K as it is a standard benchmark. For NLP, depending on the downstream task we will look at the F1-score (Question Answering), accuracy on both the macthed and unmatched dev sets (Natural Language Inference) and the unweighted average of Pearson and Spearman correlation or accuracy (Sentiment Analysis).

References

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

[1] https://github.com/pytorch/fairseq/tree/main/examples/data2vec
[2] https://cocodataset.org/

[2] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

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

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[8] 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.

[9] 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.

[10] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, 2018.