HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS
–
Fakultät IV
Wirtschaft und
Informatik

Exposé

Tim Cares

Master-Arbeit im Studiengang "'Angewandte Informatik"'

December 29, 2023

Autor Vorname Name
Matrikelnummer
dauerhafte email-Adresse

Erstprüferin: Prof. Dr. Vorname Name
Abteilung Informatik, Fakultät IV
Hochschule Hannover
email-Adresse

Zweitprüfer: Prof. Dr. Vorname Name
Abteilung Informatik, Fakultät IV
Hochschule Hannover
email-Adresse

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich das eingereichte Exposé selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Hannover, den December 29, 2023 Unterschrift

Contents

0.1 Introduction

Ever since the development of the Backpropagation algorithm Supervised Learning has been the most fundamental technique for training Deep Learning Models. This is not surprising, as it provides a model with clear domain specific learning signals, which is the most direct and efficient way of solving a problem or learning a task, respectively [1].

However, the philosophy of developing artificial intelligences using labeled data has its limits. In order for Machine Learning models, and especially Deep Learning models, to learn increasingly complex tasks requires increasingly more (labeled) data. Obviously, creating millions of labeled examples is both costly and becomes more infeasible the more complex the underlying task gets.

Because of this, Self-Supervised Learning has received an increased attention from the scientific community over the last few years. This is because Self-Supervised Learning does not rely on creating labeled data by hand, e.g. through human annotation, but receives it from the context of the data. To train a Large-Language Model (LLM) to understand written text, for example, words in example sentences are masked or deleted, respectively. The task of the model is then to predict those missing words.

The moon shines bright at night. → The [MASK] shines [MASK] at night.

Figure 0.1: An example for creating labels out of raw data. Any word in a sentence can potentially be masked, which is why training examples (including labels) can be created just from the data itself, without any human annotation (adapted from [2]).

This has three advantages: Firstly, labels do not need to be created by hand, as it is easy to randomly mask words in a sentence and use them as the targets to predict during training. Secondly, because there are massive amounts of text available on the internet, a massive amount of training data can be generated. And lastly but most importantly, the model learns to write text that represents the world we live in. This becomes clear with the example seen in Figure 0.1. Here the model would have to predict the words ”moon” and ”bright” based on the context/words remaining after masking. In order to do so successfully, the model has to learn that only the moon shines at night, not the sun, and that if the moon shines, it is usually bright.

The aforementioned example illustrates an important characteristic of Self-Supervised Learning: It forces the model to learn common sense and the world that we humans live in [2]. In some setting like vision and sound, or even natural language, this is characterized by creating a representation of the respective input, similar to word embeddings. So for an image the network might produce a vector that represents the semantic content of that image [3][4][5], which is called Representation Learning.

This makes Self-Supervised Learning suitable as a generic pre-training task: If a model can understand text, images, or sound that reflects the world we live in, then it can be fine-tuned to various downstream tasks. A model that has learned to extract the content of an image can be used to detect cats and dogs with only little fine-tuning, which is significantly simpler than training a model from scratch. It follows that for supervised tasks such pre-trained models needs less human annotated examples because they already understand the data itself.

Up until now only few models have been developed that combine Self-Supervised Learning of different modalities, most prominently natural language, vision and sound, into one single model. This by some called ”big convergence” [7] has its foundation in the fact that concepts of the real world are not bound to a specific modality, but rather expressed in one. Continuing the previous example (Figure 0.1): The concept of the moon shining at night does not change when expressed as text, photographed in an image, or spoken using sound. Therefore, the same concept should always have the same representation regardless of the modality in which the network receives it.

In order for this to work, a model is required that can process multiple modalities and create representations for the concepts that are expressed using those modalities, which is called Multimodal Representation Learning.

0.2 Objectives and Research Question

The research in the master thesis focuses on Multimodal Representation Learning and will consist of three parts.

At first, the goal is to construct a Multimodal model based on the learning tasks developed in the paper Data2Vec [8]. Joined with those efforts, the thesis will evaluate the effect of the model size on the quality of the representations, as there has been a trend to scale language and vision models to billions of parameters, making it infeasible to train them outside of big corporations. The paper will test different architecture to evaluate which provides the best trade-off between size, training speed and performance. Regarding performance, it will especially be interesting if a pre-trained multimodal model is able to achieve the same performance on downstream tasks (e.g. ImageNet-1k) as an unimodal model of the same size would.

The second part or the research will investigate the properties of representations generated by a multimodal model. This includes a thorough analysis on if the representations are modality-invariant, so if, for example, the text ”The moon shines bright at night” and a corresponding image of the moon at night will have similar, or even the same, representations.

The last part will include an investigation of latent-space arithmetic for Representation Learning. The goal here is to create representations with which one can perform arithmetic similar to word embeddings. This includes for example: Creating a representation of an image with a persons face, and creating a representation of a sentence about sunglasses. If one now adds the representation of the sunglasses to the representation of the face, and generates text from the resulting representation, then the resulting text should be about a person wearing sunglasses. In order for this to be successful, it might be necessary to develop a Multimodal Autoencoder that is both able to create representations of its input data, and to generate new data from those representations.

In summary, the contributions and research questions of the master thesis are as follows:

- Construction of a Multimodal Model based on Data2Vec.
- How do smaller models impact Representation Learning?
- Do multimodal representations match across modalities?
- Is latent-space arithmetic possible across modalities?

0.3 Methodology

0.4 Solution Ideas and Contributions

0.5 Preliminary Structure

1. Introduction
   a) Motivation
   b) Research Questions and Contributions
   c) Structure
2. Representation Learning
3. Multimodal Learning
   a) Pre-training Tasks and Requirements
   b) Data2Vec

   c) VLMo
   d) BEiT
   e) FLAVA
4. Methodology
   a) Relevance of Uncurated Datasets
   b) Datasets
   c) Metrics and Benchmarks
5. Research
   a) Experiments on Multimodal Data2Vec
   b) Study on Multimodal Latent Space
6. Analysis and Resulting Behavior
7. Outlook
8. Conclusion

0.6 Timeline

0.7 Bibliography

Bibliography

[1] Aurelien Geron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., Sebastopol, 2 edition, 2019.

[2] Yann LeCun and Ishan Misra. Meta AI. (n.d.). Self-Supervised Learning: The Dark Matter of Intelligence. AI Meta Blog. Retrieved December 29, 2023, from https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

[3] Z. Wu, Y. Xiong, S. X. Yu and D. Lin, "Unsupervised Feature Learning via Nonparametric Instance Discrimination," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 3733-3742, doi: 10.1109/CVPR.2018.00393.

[4] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9726-9735, doi: 10.1109/CVPR42600.2020.00975.

[5] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 15745-15753, doi: 10.1109/CVPR46437.2021.01549.

[6] LING, Shaoshi, et al. Deep contextualized acoustic representations for semi-supervised speech recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. S. 6429-6433.

[7] Wenhui Wang et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19175-19186

[8] Baevski, Alexei, et al. Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning. PMLR, 2022. S. 1298-1312.