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

Exposé

Tim Cares

Exposé for a Master-Thesis for the degree ‘Master Angewandte Informatik’

January 1, 2024

Author Tim Cares
1717249
tim.cares@stud.hs-hannover.de

First examiner: Prof. Dr. Volker Ahlers
Abteilung Informatik, Fakultät IV
Hochschule Hannover
volker.ahlers@hs-hannover.de

Statement of independence

I hereby declare that I have written the submitted exposé independently and without outside help, that I have not used any sources or aids other than those specified by me and that I have content taken from other works marked as such.

Hannover, January 1, 2024 Signature

Contents

1 Introduction 4

2 Objectives and Research Question 6

3 Methodology 8

4 Solution Ideas 9

5 Preliminary Structure 10

6 Timeline 11

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, increasingly more (labeled) data is required. 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 1.1: 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. The labels for this example would be ”moon” and ”bright” (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 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]. This idea is directly related to another learning task called Representation Learning.

This learning task is characterized by creating a representation of the respective input, which can be understood as a kind of interpretation of the concept that is represented using the input modality. So for an image the network might produce a vector that represents the semantic content of that image [3][4][5]. Other common modalities for which representations are also commonly created are text or even sound.

Combining the idea of Representation Learning with Self-Supervised Learning, namely masking some parts of the input and using the masked parts as a prediction target, we get a generic pre-trained model: If a model can understand text, images, or sound that reflect the world that we live in, then that model can be fine-tuned to various downstream tasks. For example, 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 Representation 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. To stay with the example given above (Figure 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, or in which it is generally expressed, respectively.

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.

2 Objectives and Research Question

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

At first, the goal is to construct a Multimodal model based on the paper Data2Vec [8], which lays the foundation to combine different modalities into one model. Because Data2Vec only provides the learning tasks to train one model jointly on text, vision and sound, the thesis will test different architectures motivated on three other papers, which specifically address Multimodal Representation Learning, namely VLMo [9], BEiT [7] and FLAVA [10].

Joined with this effort, 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 big corporations. Therefore, the thesis will test which procedure provides the best trade-off between size, training speed and performance. Regarding performance, it will 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 (one that has learned only e.g. image representation) 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 a modality-invariance of the representations, 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 of the thesis focuses on 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?
- Does latent-space arithmetic work between representations created from different modalities?

3 Methodology

The research will start with the selection of appropriate datasets, some of which will be used to train the multimodal models, while others will be used for fine-tuning the pre-trained models to receive benchmarks on which the models developed can be compared with others of the scientific community.

At first, the thesis aims to develop a multimodal Data2Vec model, as this is necessary to examine the properties of the produced representations in the second part of the thesis, and to compare them with the representations produced by a multimodal Variational Autoencoder. Another reason behind this order is to first get hands-on experience with Multimodal Representation Learning, as the examination of a multimodal latent space, using a Variational Autoencoder, has seen considerable less attention from the scientific community, which induces more risk regarding the success of such a model.

4 Solution Ideas

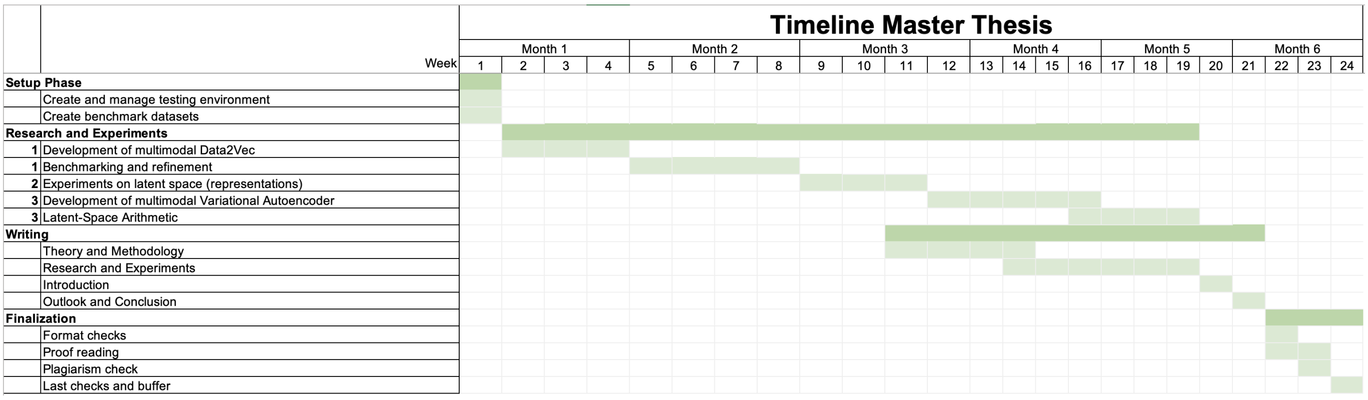
As may have become apparent in the previous chapters, the development will be based on Data2Vec [8] for the training tasks, and the architecture will be oriented on VLMo [9], BEiT [7] and FLAVA [10]. Because all models developed will be multimodal, it is inevitable that the general architecture will be based on the (Vision [12]) Transformer [11].

To evaluate the success of the latent-space arithmetic, covered in the last section of the thesis, it is necessary to develop a model which is not only able to create semantic representation of its inputs, but also able to produce new samples, namely text, images and sound, from the latent space in which the representations exist. For this a Variational Autoencoder is suitable, which is usually trained as an unimodal model, but will be adapted as a multimodal model in this thesis.

5 Preliminary Structure

1. Introduction
    a) Motivation
    b) Research Questions and Contributions
    c) Structure
2. Representation Learning
    a) Latent-Space Arithmetic
3. Multimodal Learning
    a) Pre-training Tasks and Requirements
    b) Data2Vec
    c) VLMo
    d) BEiT
    e) FLAVA
    f) Variational Autoencoders
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
    c) Latent-Space Arithmetic with Multimodal Variational Autoencoders
6. Outlook
7. Conclusion

6 Timeline



Timeline Master Thesis
Month 1
Month 2
Month 3
Month 4
Month 5
Month 6
Week
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Setup Phase
Create and manage testing environment
Create benchmark datasets
Research and Experiments
1 Development of multimodal Data2Vec
1 Benchmarking and refinement
2 Experiments on latent space (representations)
3 Development of multimodal Variational Autoencoder
3 Latent-Space Arithmetic
Writing
Theory and Methodology
Research and Experiments
Introduction
Outlook and Conclusion
Finalization
Format checks
Proof reading
Plagiarism check
Last checks and buffer

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 Non-parametric 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.

[9] Bao, Hangbo, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 2022, 35. Jg., S. 32897-32912.

[10] Singh, Amanpreet, et al. Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. S. 15638-15650.

[11] Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30. Jg.

[12] Dosovitskiy, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.