Research Questions

(Our contributions:)

# Construction of a Multimodal Model based on Data2Vec

# How do smaller models impact Representation Learning?

# Has a Multimodal Latent Space the same properties as a Unimodal Latent Space?

(Requires construction of Cross-Modal Variational Autoencoder)

# Do the representations match across modalities?

How does it perform compared to specialized models (single-modal)?

Can it be used as a general go-to pre-trained model to fine-tune to different downstream task? (Like EfficientNet for images, but for all tasks?) -> Is it a good transfer learner?

How can we make training faster while keeping performance the same/similar?

Can we go smaller? -> A step away from billion-sized models as feature extractors

Train one model based on data2vec using datasets used for data2vec training

Train the same model only using uncurated datasets (if possible) and then compare

Multi-Modal Representation Learning on Uncurated Datasets

Multi-modal ViT (Autoencoder) for representation learning and latent-space arithmetic

* latent-space arithmetic possible without Variational (AE)?

Use MoME -> Model size increases, but important is training speed, if later used for image fine tuning all experts except for vision can be discarded

How to perform fusion? Is it impossible with uncurated datasets (so no e.g. image-text pairs)

* For vision-text input: I\_CLS, Patch Embeddings, T\_CLS, Word embeddings (?)
* Vision: I\_CLS, Patch Embeddings
* Text: T\_CLS, Word embeddings
* …

Compromise by using a lot of uncurated data -> For unimodal tasks

And fewer, but still enough, curated data -> For multimodal task e.g. image-text

* Does a model with the same amount of curated data for multimodal tasks have increased performance on multimodal tasks if it has increased amounts of data for unimodal tasks?

Shared latent space for image, text and sound?

Do the representations match across modalities?

How does it compare to unimodal? Results similar?

Are they still aligned after corruption/augmentation?

* Not word embeddings, but (whole) “sentence embeddings”

Use ViT Autoencoder for continuous (image, audio) and only the encoder (+ classification head) for text?

Structure

1. Introduction
2. Representation Learning
3. Multimodal Learning
   1. Alignement of input data
   2. Pre-training tasks and requirements (Stichwort uncurated/curated(human annotated)
   3. Data2Vec
   4. VLMo (They do not have any task where they mask parts of the image, why?)
   5. BEIT
   6. FLAVA
4. Methodology
   1. Comparing models (compare with known models, and with single-modal models of the same size+architecture)
   2. Datasets used
   3. Metrics (Benchmarks)
5. Resarch/Experiments
   1. Choosing the right architecture (prob. ViT)
   2. Teacher-Student ViT
   3. ViT Autoencoder
   4. Order of training examples (vision/text/speech first, or mixed from beginning?)
   5. Analysis of/Avioding Collapse
   6. Study on different training tasks (Loss-functions)
   7. Performance after fine-tuning (benchmarking to compare with e.g. efficient-net?)
   8. Uncurated vs. curated dataset
   9. Effect of different Masking and data augmentation
   10. Only multi-modal examples vs. multi-modal and single-modal examples
   11. Single task vs. multi-task (VLMo)
6. Analysis and resulting behavior (What can the network do? What not? Maybe already in 5.?)
   1. Cross-modal Retrieval/Query similarity (explained below)
   2. Latent Space Arithmetic (inter- and intra-modal)
   3. Transitivity between modalities (synchronization -> <https://arxiv.org/pdf/1706.00932.pdf> end of chapter 2)
   4. Suitability as feature extractor (or fine-tuner) (… on multi-modal or single-modal tasks)
      1. Fine-tuning
      2. Few-shot
      3. (Low-shot)
      4. Zero-shot
7. Outlook (maybe try models of different sizes and different numbers of training examples to estimate how results would behave at extreme scale?)
8. Conclusion

## Retrieval similarity

Pass a lot of data through the network during inference (text, image (,speech)) and record their representation in higher layers or of last hidden layer.

Pass example text, image or sound through network and find top-n most similar representations recorded before.

Visualize!

Augment same input (image probably best), and do again. What changes in top-n?

(=> Basically the same seen in the Deep Learning lecture!)

-> Do the exact same but with one of the single-modal models!

-> E.g. use image model and multi-modal model, pass a lot of images (exactly the same) through networks and record, then take query images and retrieve most similar outputs. Are the retrieved images similar between both networks?

Use <https://www.researchgate.net/publication/221573542_A_New_Approach_to_Cross-Modal_Multimedia_Retrieval> ? A New Approach to Cross-Modal Multimedia Retrieval