1. Compare Self-Supervised approaches
   1. Simple convolutional Autoencoder
   2. Joint-Embedding Architecture (Siamese)
   3. Joint-Embedding Predictive Architecture
   4. Reconstruction
   5. Generative
2. Why does collapse happen in Siamese Networks? How to avoid it?
3. ViT (Autoencoders)
   1. Problem: Too big
4. Improve existing methods (+ smaller models?)
5. Multimodal pre-training using Data2Vec (Vision, Text, Speech)
   1. Better than one for each?

Multimodal Representation Learning

* Is for producing a pre-trained model that can be applied/fine-tuned for image, text (, audio)
  + Work is not done to focus on a chat bot or similar (like .e.g Google Gemini, although some parts might be interesting to learn from it, or to mention in the thesis)
  + Can a multi-modal model be fine-tuned as good (so with better or same performance) than a single modal model?
  + What about representation itself? Are they similar for fitting text and image?
    - What if one augments the image a little? Still similar?
* Train 3 separate ones
  + Or just 2 -> Exclude audio?
* Train one join
* Test performance on imagenet
  + Get reproduction of benchmark paper, e.g. SimSiam
* Train on unbalanced and uncurated
* Compare Representation of image and corresponding test
* Test how to explicitly ensure that image and fitting text have similar representations
  + Modify loss or something
  + Use Text-to-image (and vice versa) dataset
  + Also try audio
* Visualize network (Transformer) -> Explainable AI
* One can argue for multi-modal with the two sources mentioned by the following paper in the right side of page one (Introduction) -> <https://arxiv.org/pdf/1706.00932.pdf> (Can transfer between image+text and image+sound! Is this knowledge usable?)
* Datasets useful are also noted in: <https://arxiv.org/pdf/1706.00932.pdf>
* Problem regarding uncruated dataset <-> paired dataset to learn across modalities (e.g. text and image!)
* Use student-teacher that share upper layers?

Important:

* Dataset should be uncurated (+ class imbalanced)
* (Results should be a fast learner (low-shot)) -> Probably not feasible with network of low size