# Datasets and Considerations – 21.02.2024

* When selecting datasets there are obvious sources from different modalities, e.g imagenet for images.
* Which to select also depends on the learning task I guess
  + Regarding Data2Vec: If the goal is to create a model which can be used as a pre-trained model to fine-tune it to downstream tasks, then one does not necessarily need multimodal datasets.
  + So if one wants e.g. a model which can be fine-tuned on a vision-only, audio-only, text-only task, then simply training the model on different unimodal datasets should work
  + Would be modality-independent single-modal applicable model.
  + Will be interesting to see if performance on benchmark datasets improves if model trained on additional multimodal tasks (e.g. text-image)
    - Does is learn something/more over a single modality which it can then apply to this single modality (text, vision or audio) to improve on the respective single modality?
  + But one big goal is of course the alignment of representations from the same concepts given as different modalities as input.
  + So for that multimodal datasets strictly necessary!
  + Interestingly, I think this means the data has to be uncurated, so human “preprocessed” so that e.g. the text and image of a training example match
    - Unless of course one would somehow extract images from the web including their captions or alt-texts
  + So truly uncurated training in means of “true” self-supervised” learning only possible with training of unimodal datasets -> Sound, images, and text can automatically be collected, without any human supervision.
* I also have to consider how, when training on multimodal and unimodal examples jointly, how to sample the examples, FLAVA uses empirical sampling (<https://arxiv.org/pdf/2112.04482.pdf> )

## Datasets

* Use Mask-then-predict task, and no multi pretraining task, and no contrastive task(?) -> BEiT-3 says lower batch size needed (<https://arxiv.org/pdf/2208.10442.pdf> )

Training:

* ImageNet Train
* Book Corpus (It seems it is not available anymore)
* English Wikipedia (In both Data2Vec and Data2Vec 2.0 not explicitly stated which exact dataset)
  + Viable options are either Hugging Face Wikipedia, or Pytorch WikiText-2/103
  + Should be enough for this case, as I have less computation power available
  + + Another option is 15k Gutenberg Books from Kaggle (<https://www.kaggle.com/datasets/mateibejan/15000-gutenberg-books?select=guteberg_download.py>)
* Pytorch Librispeech
* Visual Genome
* CoCo Image captioning
* *Google’s Conceptual Captions (Is uncurated, as images and their html alt-text are crawled from the web)*
* (FLAVA has additional Vision-Language Datasets…)
* “See. Head. And Read: Deep Aligned Representations” use only Text-Image, Image-Audio Datasets, and show that the model automatically transfers to Text-Audio (https://arxiv.org/pdf/1706.00932.pdf)
* But Image-Text, Text-Audio seems more intuitive -> Gigaspeech of Hugging Face (<https://huggingface.co/datasets/speechcolab/gigaspeech> )
  + Also, only once image data, makes learning easier? Maybe check that
  + Train once on Image-Text + Text-Sound, and Image-Text + Image-Sound
* VALOR-1M/32k Dataset (<https://casia-iva-group.github.io/projects/VALOR/download.html>) from Paper

Testing/Evaluation:

* At least one fine-tuning test for each modality

Image

* ImageNet Top-1 Validation Accuracy
* SETI Breakthrough Listen (fine-tuning and performance on submission)

Language

* NLVR2 (Fine-tuning needed?)
* MNLI (Fine-tuning)
* SST(2) (Sentiment analysis) (Fine-tuning)

Audio

* Word error rate Librispeech test-other set

Vision-Language

* Visual Genome test set
* COCO Karpathy test set
* COCO zero shot (FLAVA)
* VQAv2 test-dev, test-std

Text-Audio

* Gigaspeech test set

Vision-Audio

* <https://www.kaggle.com/datasets/birdy654/scene-classification-images-and-audio>

General

* (General test sets to the training sets, on which the model was trained)
* Embedding similarity between examples of multimodal datasets (test set)
* (Cross-modal retrieval <https://arxiv.org/pdf/1706.00932.pdf> Fig. 4 -> Like the one visualizations method in the Deep Learning Lecture? Finding the samples with a representation closest to the one we are examining?)
* Take multiple examples from each multimodal dataset (test set), and compare representations produced by Data2Vec model trained on (1) only unimodal and (2) multimodal (+unimodal) datasets
  + Do the representations get closer?