Data Collection and Preparation

The data we need to collect has to be both unimodal and multimodal. The requirement for multimodal data is obvious: We aim to align image and text, which requires a dataset of image-text pairs. Unimodal data is required for preliminary tests of classic unimodal knowledge distillation, on which we can then build. Further, we will utilize unimodal data for the evaluation of multimodal models on downstream tasks. After all, a multimodal model should not only excel in aligning modalities, but also in tasks that only involve on of the aligned modalities. This also gives us the opportunity to compare the performance of unimodal and multimodal distilled models on the same tasks.

Unimodal Data

Collecting unimodal data does not pose an obstacle, as there are many highly curated and large datasets available. For image data, we select ImageNet-1K [1], which is an intuitive choice, as is features a high variety of content, is widely used for image classification, and, with 1.2 million training images, can be considered a medium-sized dataset. For comparison, current (August 2024) SOTA vision-language models have been trained on datasets spanning at least 14 million samples [2]–[4].

We will use this dataset for both knowledge distillation, and, most importantly, for the evaluation of image models using the ImageNet-1K validation accuracy metric, which is the most popular benchmark for computer vision models by far. We utilize the full dataset of the 2012 version, which contains 1.2 million images for training and 50,000 for validation [1]. The data can be downloaded from Huggingface's dataset hub[1] without any costs, merely requiring an account.

For raw text data, used in unimodal knowledge distillation of our text model, we select OpenWebText (OWT) [5]. This data was developed to replicate the datasets used to train GPT-2, and is also publicly available on HuggingFace[2]. The dataset consists of raw unstructured text, without any labels, which are not necessary for our distillation process. It is published as 21 chunks, and we select the first 6 for training and the 7th for validation, which is around 33% of the data. We do not collect the full dataset, as the training data, when slicing it into sections of 192 tokens, which each slice being one training example, already consists of more than 2.5 billion tokens, which we consider sufficient for our purposes. Even though the data is already preprocessed and cleaned, we further preprocess it by removing empty lines and null bytes, which we found to be quite common and lead to problems during encoding and training, as they provide no learnable information.

For benchmarking language models, including our multimodal models, on downstream tasks, we will use the GLUE benchmark [6]. GLUE, short for **G**eneral **L**anguage **U**nderstanding **E**valuation, is a collection of NLP datasets spanning four different tasks: Sentiment analysis (SST-2), grammar error detection (CoLA), sentence similarity (STS-B, MRPC, QQP), and natural language understanding (QNLI, MNLI, RTE). All 8 datasets are publicly available, and can also be accessed through HuggingFace[3].

The 8 datasets measure the performance of language models on the following tasks:

SST-2

Sentence classification of rotten tomatoes movie reviews into "negative" (1), "somewhat negative" (2), "somewhat positive" (3), and "positive" (4) [7].

---

[1]https://huggingface.co/datasets/ILSVRC/imagenet-1k
[2]https://huggingface.co/datasets/Skylion007/openwebtext
[3]https://huggingface.co/datasets/nyu-mll/glue

CoLA

Is a binary classification tasks to test a models understanding of grammar: Model should output whether a sentence is grammatically correct (label: "acceptable" -> 1) or not (label: "unacceptable" -> 0) [8].

STS-B

A regression task. The model is tasked with predicting the similarity between two sentences. The similarity score is in the interval $[0, 5] \subset \mathbb{R}$ [9].

MRPC

Is a binary classification taks. The training objective is paraphrase detection, meaning whether two sentences describe the same semantic concept [10].

QQP

The same as MRPC, instead of a simple sentence pair, the goal is to detect wethere two questions are semantic duplicates, i.e. ask the same thing[4].

QNLI

A binary classification task, where the model has to predict whether one sentence is the answer to a question represented by another sentence. Examples are of the form (question, sentence) [11], [6].

RTE

A dataset of text pairs, where the model has to predict whether a hypothesis (sentence 2) can be inferred from a text (sentence 1) [12]–[15], [6]. The task is binary classification (hypothesis can be inferred, or not).

MNLI

A classification task, where the model has to predict whether a hypothesis can be inferred from the premise (entailment), contradicts the premise (contradiction), or is neutral (neutral). There a two versions available, MNLI matched and MNLI mismatched. Both consists of the same training dataset, but test set of MNLI mismatched consists of out-of-domain data, so sentence pairs about concepts not seen during training. It is therefore a better measure of generalization, compared to MNLI matched [16].

Concrete examples can be found in (TODO: cite glue examples) in the Appendix.

Dataset	Training Examples
ImageNet-1K [1]	1.28M
OpenWebText (subset) [5]	13M
GLUE [6]	990K (total)
Total	15.27M

Table 1: Unimodal datasets and their sizes used in this work.

Multimodal Data

For training multimodal models, specifically image-text models, datasets containing image-text pairs are required. We orient ourselves on the BEiTv3 paper, currently achieving SOTA performance on multimodal benchmarks [3]. The paper uses the datasets COCO [17], Visual Genome [18], Conceptual Captions 3M [19] and 12M [20], and SBU captions [21], of which we select COCO, being

---

[4]https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

the most popular and widely used dataset and therefore an intuitive choice, and a subset of both Conceptual Captions 3M and 12M.

While the COCO dataset can be downloaded in its entirety from the COCO website[5], both variants of Conceptual Captions, developed by Google, only provide urls and the caption for each image. This is because the images used come from a variety of sources on the internet, and have been uploaded by humans all over the world. The images stem from blog posts, news articles, social media, and other sources. Since Google does not own the rights to the images, they cannot provide them in a dedicated dataset, which is why there is no guarantee that all images will be available at the time of download. Because of this, we have to utilize not only Conceptual Captions 3M, but also the 12M variant, featuring less curated data [19], [20]. The index of CC3M is available on the official Conceptual Captions website (training split)[6], and the index of CC12M can be found in the corresponding GitHub repository[7].

Another popular choice for image-text pairs is the Visual Genome dataset [18], containing high quality images and detailed annotations. However, we refrain from using this dataset, as we experienced unstable training during preliminary tests, a circumstance we will address again in the experimental part of this work.

Dataset	Avg. Caption Length	$\frac{\text{\# Captions}}{\text{\# Images}}$	# Images	# Image-Text Pairs
COCO [17]	11.0	5.0	82,783	566,747
CC3M (subset) [19]	12.0	1.0	1,516,133	1,516,133
CC12M (subset) [20]	10.3	1.0	1,181,988	1,181,988
Total	-	-	2,780,904	3,264,868

Table 2: Multimodal Dataset used for aligning image and text.

On Curated Datasets

The goal of this work is to develop a multimodal model that is cheap to train, and does not rely on labeled data in the end-to-end process. That means not only should our multimodal model not require labeled data for training, but also any pretrained models used in the process.

Whether image-text datasets, and, in fact, any multimodal dataset consisting of pairs of data, can be seen as curated or even labeled data is a matter of perspective. While the data is not labeled in the traditional sense, as in having a label for an image or text, the pairs themselves can be seen as labels. While single images or texts can be considered as in-the-wild data, i.e. data that appears naturally in the real world, like in books, articles, or on the internet, image-text pairs require image and text to be paired together. This can be seen as less natural, as it requires a human to create the caption for an image, or vice versa, which is a form of labeling. The COCO dataset, for example, can be seen as labeled, as for each caption a human created a caption with the specific intention of creating data for Machine Learning. Consequently, whether multimodal learning can be seen as self-supervised learning, as it is often referred to in the literature [3], [2], [4], is debatable. With this in mind, creating a multimodal model that is scalable in the sense that it does not rely on labeled data, which is one of the most challenging aspects of AI research, is, if multimodal data is seen as labeled data, not possible.

However, there are indeed multimodal data sources that can theoretically be seen as unlabeled. One example is the alt-text of images on the internet. Even though the alt-text is created by humans, it is

---

not created with the intention of creating data for Machine Learning, but rather to provide a description of the image for visually impaired people. Consequently, the data was generated as a byproduct of a different task, and can therefore be seen as unlabeled, or at least uncurated data.

This is exactly why we select both CC3M and CC12M, as they, especially CC12M, consists of in-the-wild image-text pairs from the internet [19], [20]. This way of collecting data and training models is therefore significantly more scalable than using curated datasets specifically created for Machine Learning.

(a) COCO Captions

A large sheep bends his head towards some grass.

(b) Conceptual Captions 12M

#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life
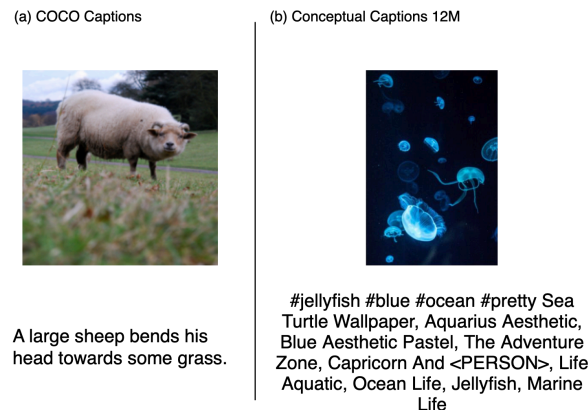
Figure 1: Side-by-side comparison of examples seen in COCO (a) and CC12M (b). While COCO features high quality images and detailed annotations, CC12M consists of in-the-wild image-text pairs from the internet, and image-text pairs therefore might be less correlated. However, the latter enables scalability, as more data can be collected without the need for human annotation. Images and text in the figure have been taken from the COCO train set [17] and CC12M [20], respectively.

Data Persistence

Data Selection and Collection:

- starting with unimodal:

- collecting unimodal data is not a problem, many highly curated and large datasets available

- because we are using Knowledge-Distillation based on self-supervised distillation, so we do not need labels for the distillation process, we can use any image dataset

- for image data, we select imagenet

  ‣ build for image classification and object detection (labled, but, again, not necessary)

  ‣ each image corresponds to one of 1k classes

  ‣ very popular, high quality, high variety -> 1000 classes, by standards of SOTA models it is a medium sized dataset (ca. 1.2M train images)
    – why medium sized? -> papers used in this thesis have been trained on much larger data -> VLMo around 14 million image(-text) examples,

    BEiT on more than 35M, and FLAVA (only mentioned a couple of times) even 70M
    – models also much larger than models build here, so we do not need as much, nor is it feasible for us to train on that much data

  ‣ Data2Vec Image model exclusively trained on imagenet

  ‣ Data2Vec, BEiT, VLMo, and FLAVA all use imagenet for evaluation -> we should use it as well

  ‣ we use the full dataset of the 2012 version, with 1.2M images for training and 50k for validation

- ‣ during pretraining, i.e. the Knowledge-Distillation, we apply the same data augmentation as in Data2Vec2
  - – as we also use Data2Vec2 as the teacher in many experiments, this allows us a close comparison between the models, as they are trained on the same data
  - – data augmentation includes random resized crop, followed by a random horizontal flip, followed by normalization each channel seperatly using channel wise mean and standard deviation computed from the imagenet training set -> standard procedure for image preprocessing, and used for all images throughout this work
  - – random resized crop: crop a random part of the image, then resize it to the desired size (224x224)
    - • crop size is hyperparameter, but we just use the same as in Data2Vec2, which is 0.08 to 1.0 of the original image size
    - • 8% as lower bound seems to be a very low value, a lot of information is lost, but it is a common value in the literature, so we do

      the same
  - – random horizontal flip: randomly flip the image horizontally (self-explanatory)

- ‣ for validation, we resize each image to the same size as in training (224x224) and normalize it using the same procedure as in training

- ‣ generally ALL images in this thesis will be scaled to the same size: 224x224

- ‣ we access the data from Huggingface's dataset hub

- benchmarks published on downstream tasks/dataset like GLUE, more on that later

- data is not that much and only meant for fine-tuning and benchmarking as downstream task

- as with images, we a just trying the replicate the outputs, i.e. the representations of the input data

- we can use any text (dataset(s)), as long as it is large enough

- for text data (pretraining, which is the Knowledge-Distillation that we do) we select openwebtext
  - ‣ dataset build to reproduce datasets used to train GPT-2
  - ‣ publicly available and popular, used by e.g. BEiT3

- we access the data from Huggingface's dataset hub
  - ‣ published as slices and without any split
  - ‣ we take subsets 0-4 for training and 5 for validation, which is about 25% of the data

- due to open source efforts, data is already preprocessed and cleaned

- we apply further preprocessing by removing empty lines and null bytes, which we found are quite common and lead to problems during encoding and training, as they provide no learnable information

- the text of every dataset, containing text, so also openwebtext, is tokenized and encoded using the GPT-2 byte-pair encoder (citation here -> same as in D2V), with a vocabulary size of 50262 tokens

- we seperate the sentences using the end-of-sentence token, also done by Data2Vec2

- also used by Data2Vec2, and we use it, again, for the purpose of comparison

- so save disk space, we save the training and validation sets of owt in a single binary file, respectively
  - ‣ the binary files already contain the encoded text, so that we only need to batch them during training

- ▸ in order to ensure correct encoding and to save the time for implementing the binary encoding, we use the dataset functionality of Fairseq, a library for sequence-to-sequence models developed by Facebook/Meta, to encode an binarize the text data, which is also used by Data2Vec2

- in total, our openwebtext subset consists of more than 2.5 billion tokens and consumes roughly 6 GB of disk space

- really low, compared to the image data, which is about 150 GB

- we do not use bookcorpus and english wikipedia, datasets Data2Vec2 was trained on, as Openwebtext appears to be a more recent and popular dataset for Knowledge-Distillation -> DistilGPT2 and Distilroberta trained on openwebtext, results showed that this dataset yields good results for (knowledge) distillation

- multimodal data:

- we need to use datasets with image-text pairs

- as Data2Vec not multimodal, we do not have any reference datasets

- however, many multimodal models, like BEiT and VLMo use the same popular multimodal dataset

- we therefore also opt for them

- we use COCO and a subset of Google's Conceptual Captions

- even though COCO contains just contains 82783 images, which is not that much, it contains multiple captions per image, meaning we can create multiple image-text pairs from one image

- images have a little more than average 5 captions, yielding a total of 566747 actual examples for the training set (used for Knowledge-Distillation)

- we also additionally use SBU Captions (SBU) and a subset of Google's Conceptual Captions (CC3M), which originally contain 1M and over 3.3M unique image-text pairs, respectively

- COCO has comparatively few images, just 82,783, so we need more data to train the model

- also helps in balancing the ratio of unique images and text

- with just COCO, we have five times more variety in text than in images

- in SBU and CC3M each image is only associated with one caption, and because we use at least 10x more images than in COCO, we can reduce the relative overrepresentation of text

- both datsets do not provide the images directly, but instead an index with the url of an image, and a correspondig caption

- url point to various sources on the web, so there is no guarantee that all images are still available

- for SBU, we collect all image(-text pairs) that are available as of July 2024

- for Google's CC3m we use the first 800,000 available, as of July 2024, images (with their captions) as a subset from the training set index published by Google[8].

- storing the whole dataset is not feasible for us and the combination with COCO and SBU should already provide enough data for training, as shown in table Table 2
  - ▸ we are constrained by 600GB of disk space, which must also be shared with other datsets, e.g. ImageNet-1k

- ids and urls of image-text pairs used, for both SBU and CC3M, are available on GitHub *< TODO: Footnote with URL!!! >*

- aforementioned papers (e.g. BEiT-3) also use the popular Visual Genome (VG) dataset, containing 108,249 images with on average 50 captions per image

- we have that many captions, because the captions are actually region descriptions of the images, not a general/global description of the image [18]
  ‣ can be very short and only capture a small part of the image
  ‣ are used for describing regions/objects in the image

- still used by the papers

- we will not use it as we encountered problems when using it together with contrastive learning, which will be explained in section

- all captions are tokenized and encoded using the same GPT-2 byte-pair encoder as the text-only data

- as usual, and done in BEiT, VLMo, and FLAVA, we prepend each caption with a start-of-sequence token and append an end-of-sequence token

- we use the same data augmentation as in the unimodal case for the images
  ‣ that is, during training we apply random resized crop, followed by a random horizontal flip, followed by the imagenet normalization
  ‣ during validation, we resize the image to 224x224 and normalize it using the same procedure as in training
  ‣ the only difference lies in the crop size of the images
  ‣ min crop size set to 0.08 for image pretraining (unimodal image distillation)
  ‣ means at a minimum we could crop 8% of the image, and discard the rest
  ‣ destroys a lot of information, not a problem for image only training, but when the image has a caption describing the image, and some parts focusing on the cropped parts, which can especially happen for VG, where the captions consists of region descriptions, then we might have captions that do not match the image anymore
  ‣ that is why papers that use random crop use higher values -> BEiT3 uses 0.5, FLAVA 0.9, VLMo uses RandAugment
  ‣ we consider 0.5 too low, and because our model is smaller, and we use less data, we are more prone to noise, so we use FLAVA's value of 0.9
  ‣ danger is high that semantic content, described in the caption, is lost

- examples of image-text pairs and the effect of the crop size can be seen in the appendix

- in order to evaluate the models performance on text-only tasks, we use the GLUE benchmark

- GLUE consists of 4 different tasks: sentiment analysis (SST-2), grammar error detection (CoLA), sentence similarity (STS-B, MRPC, QQP), and natural language understanding (QNLI, MNLI, RTE)

- single questions encoded using same GPT-2 byte-pair encoder as before, and prepended with start-of-sequence token and appended with end-of-sequence token

- sentences are tokenized (samte tokenizer), concatenated and separated by the end-of-sentence token, concatenated sentence pair prepended with start-of-sequence token

---

[8]https://ai.google.com/research/ConceptualCaptions/download

Dataset	Example	Label
CoLA	Our friends won't buy this analysis, let alone the next one we propose.	1
SST-2	hide new secretions from the parental units	0
MRPC	Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. [SEP] Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.	1
STS-B	A plane is taking off. [SEP] An air plane is taking off.	5.0
QQP	How is the life of a math student? Could you describe your own experiences? [SEP] Which level of prepration is enough for the exam jlpt5?	0
MNLI	Conceptually cream skimming has two basic dimensions - product and geography. [SEP] Product and geography are what make cream skimming work.	1
QNLI	When did the third Digimon series begin? [SEP] Unlike the two seasons before it and most of the seasons that followed, Digimon Tamers takes a darker and more realistic approach to its story featuring Digimon who do not reincarnate after their deaths and more complex character development in the original Japanese.	1
RTE	No Weapons of Mass Destruction Found in Iraq Yet. [SEP] Weapons of Mass Destruction Found in Iraq.	1

Table 3: Training examples of the GLUE benchmark tasks (one example per task).

Bibliography

[1]   O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[2]   H. Bao *et al.*, "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts," in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://openreview.net/forum?id=bydKs84JEyw

[3]   W. Wang *et al.*, "Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19175–19186. doi: 10.1109/CVPR52729.2023.01838.

[4]   A. Singh *et al.*, "FLAVA: A foundational language and vision alignment model," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2112.04482

[5]   A. Gokaslan and V. Cohen, "OpenWebText Corpus." 2019.

[6]   A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," 2019.

[7]   R. Socher *et al.*, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://www.aclweb.org/anthology/D13-1170

[8]   A. Warstadt, A. Singh, and S. R. Bowman, "Neural Network Acceptability Judgments," *arXiv preprint arXiv:1805.12471*, 2018.

[9] P. May, "Machine translated multilingual STS benchmark dataset.," 2021. [Online]. Available: https://github.com/PhilipMay/stsb-multi-mt

[10] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the International Workshop on Paraphrasing*, 2005.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of EMNLP*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 2383–2392.

[12] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment.* Springer, pp. 177–190, 2006.

[13] R. Bar Haim *et al.*, "The second PASCAL recognising textual entailment challenge," 2006.

[14] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The third PASCAL recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, 2007, pp. 1–9.

[15] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini, "The Fifth PASCAL Recognizing Textual Entailment Challenge," 2009.

[16] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in *Proceedings of NAACL-HLT*, 2018.

[17] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.

[18] R. Krishna *et al.*, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 32–73, May 2017.

[19] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, I. Gurevych and Y. Miyao, Eds., 2018, pp. 2556–2565.

[20] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts," in *CVPR*, 2021.

[21] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 1143–1151.