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

Titel der Arbeit

Tim Cares

Master's thesis in Applied Computer Science

25. September 2024

Autor Tim Cares
Matrikelnummer
Email Adresse

Erstprüfer Prof. Dr. Vorname Name
Abteilung Informatik, Fakultät IV
Hochschule Hannover
Email Adresse

Zweitprüfer Prof. Dr. Vorname Name
Abteilung Informatik, Fakultät IV
Hochschule Hannover
Email Adresse

I hereby declare that I have written and submitted this thesis independently, without any external help or use of sources and aids other than those specifically mentioned by me. I also declare that I have not taken any content from the works used without proper citation and acknowledgement.

Hannover, 25. September 2024 Tim Cares

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae.

Abstract

Note:

1. **paragraph:** What is the motivation of your thesis? Why is it interesting from a scientific point of view? Which main problem do you like to solve?
2. **paragraph:** What is the purpose of the document? What is the main content, the main contribution?
3. **paragraph:** What is your methodology? How do you proceed?

Zusammenfassung

Note: Insert the German translation of the English abstract here.

Contents

1 Example with Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae sine metu degendae praesidia firmissima. – Filium morte multavit. – Si sine causa, nollem me ab eo delectari, quod ista Platonis, Aristoteli, Theophrasti orationis ornamenta neglexerit. Nam illud quidem physici, credere aliquid esse minimum, quod profecto numquam putavisset, si a Polyaeno, familiari suo, geometrica discere maluisset quam illum etiam ipsum dedocere. Sol Democrito magnus videtur, quippe homini erudito in geometriaque perfecto, huic pedalis fortasse; tantum enim esse omnino in nostris poetis aut inertissimae segnitiae est aut fastidii delicatissimi. Mihi quidem videtur, inermis ac nudus est. Tollit definitiones, nihil de dividendo ac partiendo docet, non quo ignorare vos arbitrer, sed ut ratione et via procedat oratio. Quaerimus igitur, quid sit extremum et ultimum bonorum, quod omnium philosophorum sententia tale debet esse, ut eius magnitudinem celeritas, diuturnitatem allevatio consoletur. Ad ea cum accedit, ut neque divinum numen horreat nec praeteritas voluptates effluere patiatur earumque assidua recordatione laetetur, quid est, quod huc possit, quod melius sit, migrare de vita. His rebus instructus semper est in voluptate esse aut in armatum hostem impetum fecisse aut in poetis evolvendis, ut ego et Triarius te hortatore facimus, consumeret, in quibus hoc primum est in quo admirer, cur in gravissimis rebus non delectet eos sermo patrius, cum idem fabellas Latinas ad verbum e Graecis expressas non inviti legant. Quis enim tam inimicus paene nomini Romano est, qui Ennii Medeam aut An-

tiopam Pacuvii spernat aut reiciat, quod se isdem Euripidis fabulis delectari dicat, Latinas litteras oderit? Synephebos ego, inquit, potius Caecilii aut Andriam Terentii quam utramque Menandri legam?

2 Introduction

Note: Introduce the topic of your thesis, e.g. with a little historical overview.

List of Figures

List of Tables

2.1.1 Data and Preparation

General:

- data we need to collect has to be both unimodal and multimodal
  - multimodal obvious -> needed to align modalities, as described in e.g. section about "(See, Hear, and Read:) Deep Aligned Representations"
  - multimodal means in this case dataset of image-text pairs
  - unimodal -> needed for first tests and poc of distillation process
  - unimodal data also needed for evaluation and comparison of unimodal models from research papers, like Data2Vec, as well as comparison between unimodal and multimodal distilled models (models of this thesis)
  - unimodal data also needed for stage-wise knowledge distillation in section about "Mixing Positional Encodings" (will be elaborated on in the respective section)

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, Visual Genome, 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)
- same goes for Visual Genome: contains 108249 images
- here, each image has on average even 50 captions yielding 5408689 unique image-text pairs (examples) for training
- we have to consider why that is -> 50 captions quite a lot per image
  - the reason why that is is because captions are extracted from region descriptions of images
  - describe, not as in coco, the focus of the image, or rather a short description/summary of the image, but individual, sometimes

  small, regions of the image -> VG also used for object detection
- we also additionally use a subset of Google's Conceptual Captions, which originally contains over 3.3M unique image-text pairs
- we use this, because with COCO and VG we have a overrepresentation of text -> for each image multiple captions -> higher variety of text than images
- in CC3m each image is only associated with one caption, reducing the relative overrepresentation of text

- Google only provides an index file with captions to image urls
  - urls are sourced from the web, so there is no guarantee that all images are still available
- we use the first 500,000 available, as of June 2024, images (with their captions) as a subset from the training set index published by Google[1].
- storing the whole dataset is not feasible for us and the combination with COCO and VG should already provide enough data for training
- ids and urls of image-text pairs used are available on GitHub

Dataset	# Images	Avg. Captions	Avg. Caption Length	# Image-Text Pairs
COCO	82,783	5.0	11.0	566,747
VG	108,249	50.0	4.7	5,408,689
CC3M Subset	500,000	1.0	50.0	500,000
Total	691,032	-	-	6,475,436

Table 1: Multimodal Dataset used for aligning Image and Text. The maximum text sequence length is, inspired by BEiT3, set to 64 tokens. The concat version of VG will therefore also have a maximum text sequence length of 64 tokens.

- 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 0.9 too high and 0.5 too low, so we opt for 0.6

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

- 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)

SST-2:
- sentence classification of rotten tomatoes movie reviews into "negative" (1), "somewhat negative" (2), "somewhat positive" (3), and "positive" (4) [Soc+13]

CoLA:
- grammatical error detection / linguistic accepability, binary classification, whether a sentence is grammatically correct (acceptable -> 2) or not (0 -> unacceptable) [WSB18]

STS-B:
- similarity of two sentences, regression task, similarity score between 0 and 5 [May21]

MRPC:
- paraphrase detection, whether two sentences describe the same content/concept, binary classification [DB05]

QQP:
- do two questions ask the same thing, binary classification[2]

QNLI:
- does a text contain the answer to a question? [Raj+16, Wan+19]

RTE:
- pair of text and hypothesis, whether the hypothesis can be inferred from the text, binary classification [Ben+09, DGM06, Gia+07, Bar+06, Wan+19]

MNLI:

- pair of premise and hypothesis, whether the hypothesis can be inferred from the premise (entailment), contradicts the premise (contradiction), or is neutral (neutral) [WNB18]
- 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

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

Dataset	Example	Label
CoLA	Our friends won't buy this analysis, let alone the next one we propose.	1
MNLI	Conceptually cream skimming has two basic dimensions - product and geography. [SEP] Product and geography are what make cream skimming work.	1
CC3M Subset	500,000	

Table 2: Glue Example

2.2 Multimodal Knowledge Distillation

2.3 Seperate Self-Attention

2.3.1 Baseline

- currently only 6 layers, 5 out of which are modality specific, 1 is shared
- we experiment with adding one additional moadality specific layer, and one additional shared layer in another experiment

-> more difficult to align mutliple modalities, than just training one -> add one layer -> motivation for modality specific: after 5 layers information might not be high level enough so that one layer can process the information -> add one additional modality specific -> motivation for shared: after 5 layers information might be high level enough, but capturing modality agnostic information might take more than one layer -> add one additional shared

- added shared layer improves performance slightly, but adds 7 million parameters and 41 minutes to training time
- looking at the improvement in zero-shot, which increases the average Recall from 29.93% to 30.8%, this is not much of an improvement, considering the amount of parameters we add to the model

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
Data2Vec2	0.02	0.08	0.22	0.01	0.10	0.19	0.02	0.12	0.26	0.02	0.06	0.12
MM-D2V2 (Ours)	4.24	12.12	17.96	1.77	6.54	10.91	1.2	4.88	8.18	0.54	2.52	4.58
MM-D2V2 (Ours)†	31.72	56.78	67.9	12.42	31.05	42.5	7.7	26.18	37.6	4.08	17.01	24.26
MM-D2V2 7_2(Ours)†	32.78	58.34	69.3	12.83	31.85	43.4	8.08	27.92	38.6	4.14	17.5	24.82
MM-D2V2 7(Ours)†	30.24	56.48	67.46	11.96	30.48	41.88	7.36	26.42	36.6	3.7	16.58	23.84

Table 3: Comparison of Zero-shot Image-Text and Text-Image Retrieval of first results with FLAVA and Data2Vec2 papers. Because Data2Vec2 is a unimodal model, we embed each image with the D2V2-Image model and each text with the D2V2-Text model. This yields unusable results, as there has been no incentive for the models to learn a shared representation, as both are unimodal. This is why we had to use both the image and the text model to embed the data.

†: This version has been trained with BEiT-2 as the teacher model, not the D2V2 Image model.

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-Shot												
FLAVA	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
MM-D2V2 (Ours)	31.72	56.78	67.9	12.42	31.05	42.5	7.7	26.18	37.6	4.08	17.01	24.26
Finetune												
BEiT-3	84.8	96.5	98.3	67.2	87.7	92.8	98	100	100	90.3	98.7	99.5
VLMo	74.8	93.1	96.9	57.2	82.6	89.8	92.3	99.4	99.9	79.3	95.7	97.8

Table 4:

- looking at the validation loss of image and text seperatly, on COCO val set, we observe that the loss on images is significantly lower than the loss on text, which might be due to the fact that the teacher model is a vision model and the target, the cls token, might be biased towards the image modality, as it is unimodal
- interestingly, this bias also seem to be directly translated to the performance on image-text retrieval, as the performance on image-text retrieval is significantly higher than on text-image retrieval -> we are learning the cls token representation, and using the leared

cls token as an output for the student model, to encode a modality for retrieval and other downstream task

-> suggests that the cls token is biased towards the image modality, or rather that the model is better in encoding images than text

- we can see that the performance of e.g. BEiT-3 and VLMo is also lower on text-image retrieval than on image-text retrieval, but not the the extend that we observe with our model

Model	**MSCOCO (5K test set)**		**Flickr30K (1K test set)**	
	Image → Text	Text → Image	Image → Text	Text → Image
MM-D2V2 7(Ours)†	51.39	28.11	23.46	14.71
BEiT-3	93.2	82.57	99.33	96.17
VLMo	88.27	76.53	97.2	90.93

Table 5: Average recall of image-text and text-image retrieval on MSCOCO and Flickr30K. All models continously perform better on image-text retrieval than on text-image retrieval, but the difference is more pronounced for our model.

- currently vl layer(s) (or rather mulimodal layer(s)) are randomly initialized, one option is to specifically initialize the multimodal layers with the weight of the final layers of the D2V text model -> initial state is closer closer to text modality
- we do not compare with See, Hear, and Read: Deep Aligned Representations [AVT17], as they did not use the karpathy splits [KF15], use the average median rank instead of recall at a specific percent, and from their experimental setup it is not clear which samples they used from Visual Genome for their retrieval experiments.

A Supplementary Material Images

– Supplementary Material –

B Supplementary Material Source Code

Bibliography

[Soc+13] 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

[WSB18] A. Warstadt, A. Singh, and S. R. Bowman, "Neural Network Acceptability Judgments," *arXiv preprint arXiv:1805.12471*, 2018.

[May21] P. May, "Machine translated multilingual STS benchmark dataset.," 2021. [Online]. Available: https://github.com/PhilipMay/stsb-multi-mt

[DB05] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the International Workshop on Paraphrasing*, 2005.

[Raj+16] 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.

[Wan+19] 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.

[Ben+09] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini, "The Fifth PASCAL Recognizing Textual Entailment Challenge," 2009.

[DGM06] 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.

[Gia+07] 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.

[Bar+06] R. Bar Haim *et al.*, "The second PASCAL recognising textual entailment challenge," 2006.

[WNB18] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in *Proceedings of NAACL-HLT*, 2018.

[AVT17] Y. Aytar, C. Vondrick, and A. Torralba, "See, Hear, and Read: Deep Aligned Representations," *arXiv preprint arXiv:1706.00932*, 2017, [Online]. Available: https://arxiv.org/abs/1706.00932

[KF15] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE, 2015, pp. 3128–3137.