Multimodal Deep Learning

Contents

Preface

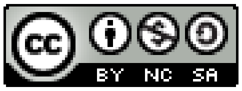

FIGURE 1: Creative Commons License

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/

Foreword

Author: Christoph Molnar

This book is the result of an experiment in university teaching. Each semester, students of the Statistics Master can choose from a selection of seminar topics. Usually, every student in the seminar chooses a scientific paper, gives a talk about the paper and summarizes it in the form of a seminar paper. The supervisors help the students, they listen to the talks, read the seminar papers, grade the work and then … hide the seminar papers away in (digital) drawers. This seemed wasteful to us, given the huge amount of effort the students usually invest in seminars. An idea was born: Why not create a book with a website as the outcome of the seminar? Something that will last at least a few years after the end of the semester. In the summer term 2019, some Statistics Master students signed up for our seminar entitled "Limitations of Interpretable Machine Learning". When they came to the kick-off meeting, they had no idea that they would write a book by the end of the semester.

We were bound by the examination rules for conducting the seminar, but otherwise we could deviate from the traditional format. We deviated in several ways:

1. Each student project is part of a book, and not an isolated seminar paper.
2. We gave challenges to the students, instead of papers. The challenge was to investigate a specific limitation of interpretable machine learning methods.
3. We designed the work to live beyond the seminar.
4. We emphasized collaboration. Students wrote some chapters in teams and reviewed each others texts.

Technical Setup

The book chapters are written in the Markdown language. The simulations, data examples and visualizations were created with R (R Core Team, 2018). To combine R-code and Markdown, we used rmarkdown. The book was compiled with the bookdown package. We collaborated using git and github. For details, head over to the book's repository[2].

[2] https://github.com/slds-lmu/seminar_multimodal_dl

1

Introduction

Author:

Supervisor:

1.1 Intro About the Seminar Topic

1.2 Outline of the Booklet

2

Chapter 1

Authors: Author 1, Author 2

Supervisor: Supervisor

2.1 Lorem Ipsum

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

R Core Team (2018)

2.2 Using Figures

Referencing can be done by using the chunk label e.g. `\@ref(fig:ch01-figure01)` for 2.1.

NOTE!!! Do not use underscores in chunk labels! This will crash the compilation ...

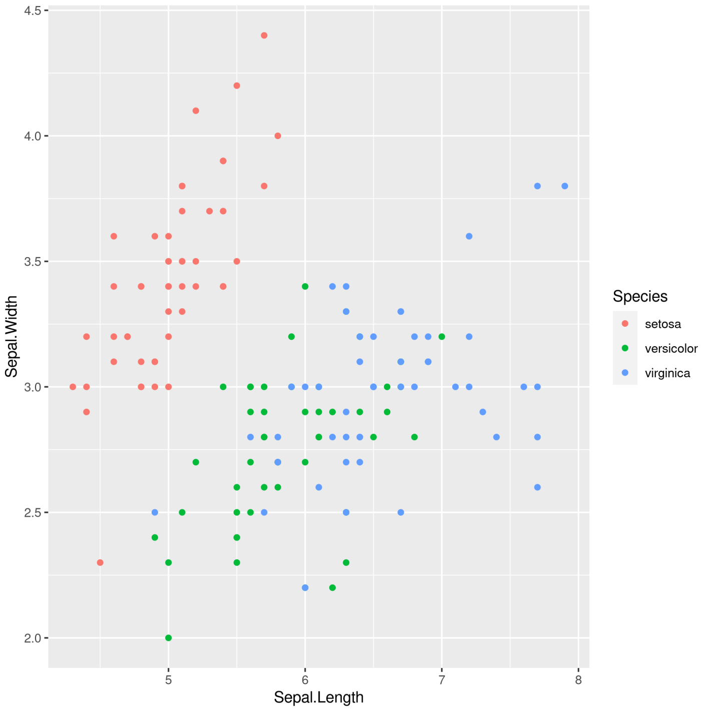

FIGURE 2.1: This is the caption of the figure!

2.3 Using Tex

HTML rendering uses MathJax while pdf rendering uses LaTeX:

$$f(x) = x^2$$

2.4 Using Stored Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1713	0.2798	7.760	0.0000
Sepal.Width	0.4959	0.0861	5.761	0.0000
Petal.Length	0.8292	0.0685	12.101	0.0000
Petal.Width	-0.3152	0.1512	-2.084	0.0389
Speciesversicolor	-0.7236	0.2402	-3.013	0.0031
Speciesvirginica	-1.0235	0.3337	-3.067	0.0026

3

title

Author:

Supervisor:

4

title

Author:

Supervisor:

5

title

Author:

Supervisor:

6

Chapter 1

Authors: Author 1, Author 2

Supervisor: Supervisor

6.1 Lorem Ipsum

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

R Core Team (2018)

6.2 Using Figures

This is the caption of the figure! This is the caption of the figure!

Referencing can be done by using the chunk label e.g. `\@ref(fig:ch01-figure01)` for 2.1.

NOTE!!! Do not use underscores in chunk labels! This will crash the compilation …

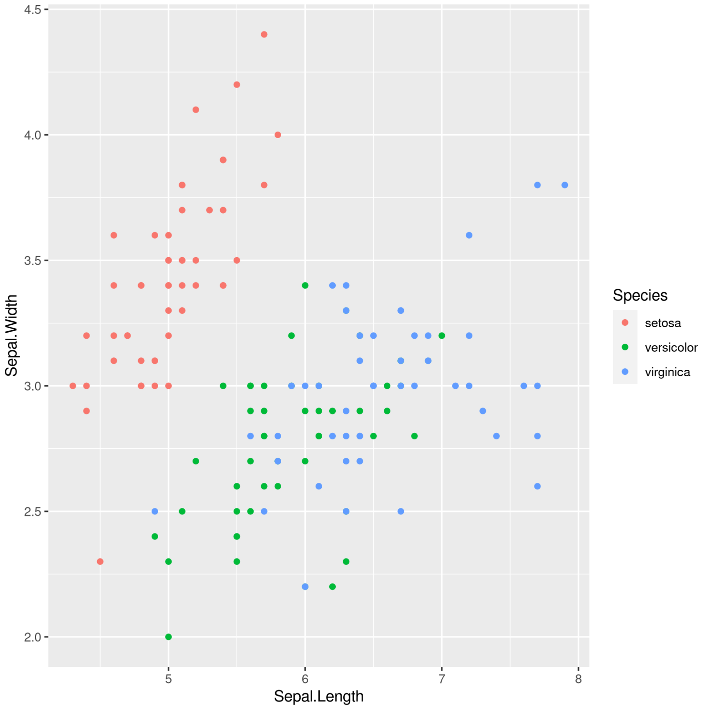

FIGURE 6.1: This is the caption of the figure!

6.3 Using Tex

HTML rendering uses MathJax while pdf rendering uses LaTeX:

$$f(x) = x^2$$

6.4 Using Stored Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1713	0.2798	7.760	0.0000
Sepal.Width	0.4959	0.0861	5.761	0.0000
Petal.Length	0.8292	0.0685	12.101	0.0000
Petal.Width	-0.3152	0.1512	-2.084	0.0389
Speciesversicolor	-0.7236	0.2402	-3.013	0.0031
Speciesvirginica	-1.0235	0.3337	-3.067	0.0026

7

title

Author:

Supervisor:

8

title

Author:

Supervisor:

9

title

Author:

Supervisor:

10

title

Author:

Supervisor:

11

title

Author:

Supervisor:

12

title

Author:

Supervisor:

13

Chapter 2 Multimodal architectures

Authors: Luyang Chu, Karol Urbanczyk, Giacomo Loss, Max Schneider, Steffen Jauch-Walser

Supervisor: Christian Heumann

13.1 Introduction

Multimodal learning refers to the process of learning representations from different types of input modalities, such as image data, text or speech. Due to methodological breakthroughs in the fields of Natural Language Processing (NLP) as well as Computer Vision (CV), in recent years multimodal models have gained increasing attention as they are able to strengthen predictions and better emulate the way humans learn. This chapter focuses on discussing images and text as input data. The remainder of the chapter is structured as follows:

The first part "Image2Text" discusses how transformer-based architectures improve meaningful captioning for complex images using a new large scale, richly annotated dataset COCO (Lin et al., 2014; Cornia et al., 2020). Whether it is seeing a photograph and describing it or parsing a complex scene and describing its context, it is not a difficult task for humans. But it is much more complex and challenging for computers. We start with focusing on images as input modalities. In 2014 Microsoft COCO was developed with a primary goal of advancing the state-of-the-art (SOTA) in object recognition by diving deeper into a broader question of scene understanding (Lin et al., 2014). COCO stands for Common Objects in Context. It addresses three core problems in scene understanding: object detection (non-iconic views), segmentation, and captioning. For tasks like machine translation and language understanding in NLP, transformer-based architecture is widely used. However, the potential of these applications in the multi-modal context has not been fully covered. With the help of the COCO dataset, a transformer-based architecture: Meshed-Memory Transformer for Image Captioning (M^2) will be introduced to improve both image encoding and the language generation steps (Cornia et al., 2020). The performance of the (M^2) Transformer and different fully-attentive models will be evaluated and compared on the COCO dataset.

Next, in “Text2Image”, the idea of incorporating textual input in order to generate visual representations is described. Current advancements in this field have been made possible largely due to recent breakthroughs in NLP, which first allowed for learning contextual representations of text. Transformer-like architectures are being used to encode the input into embedding vectors, which are later helpful in guiding the process of image generation. The chapter looks into details and discusses two SOTA model architectures by OpenAI, which both condition on text representations. Surprisingly, none of them uses a GAN approach - a method which probably has been seen as the go-to idea for image generation over the last years. The first model is DALL-E (Ramesh et al., 2021), which essentially combines Variational Encoder (VAE) with Autoregressive Transformer. In the first step, VAE is being trained to learn downsized image representations. Such embeddings are concatenated with text embeddings into one text-image pair input. However, both of them use different dimensionality and vocabulary size. In the second step, the transformer is trained on a next token prediction task given these data pairs. Finally, at inference time, the model is able to generate images in the following way:

1. Encode text input into text embedding
2. Use trained transformer from step 2 to generate image embedding
3. Use VAE from step 1 to generate image from image embedding

The next approach to text-to-image generation is a GLIDE model (Nichol et al., 2021). GLIDE stands for Guided Language to Image Diffusion for Generation and Editing. Its idea is to use Diffusion Models. In its core, Diffusion Model is a simple idea – random noise is being added to the image in an iterative fashion, and then model learns how to reconstruct this image. In the case of GLIDE this learning process is conditioned on the text prompt, which is first passed through a transformer. Both models differ in their results. While DALL-E’s resulting images might have been overwhelming back in the beginning of 2021, GLIDE is thought to significantly improve on photorealism and resolution the generated images. Since the field has already seen further improvements following GLIDE, these new developments are also going to be mentioned in the chapter.

The third part, “Images supporting Language Models”, deals with the integration of visual elements in pure textual language models. Distributional semantic models such as Word2Vec and BERT assume that the meaning of a given word or sentence can be understood by looking at how (in which context) and when the word or the sentence appear in the text corpus, namely from its “distribution” within the text. But this assumption has been historically questioned, because words and sentences must be grounded in other perceptual dimensions in order to understand their meaning (see for example the “symbol grounding problem”; Harnad, 1990). For these reasons, a broad range of models has been developed with the aim to improve pure language models, leveraging on the addition of other perceptual information, such as visual ones. This subchapter focuses in particular on the integration of visual elements (images) to support pure language models for various tasks at the word-level and sentence-level. The starting point is always a language model, on which visual representations (extracted often with the help of large pools of images like MS COCO, see chapter “Img2Text” for further references) are to be “integrated”. But how? There has been proposed a wide range of solutions: On one side of the spectrum, textual elements and visual ones are learned separately and then “combined” together whereas on the other side, the learning of textual and visual features takes place simultaneously/jointly.

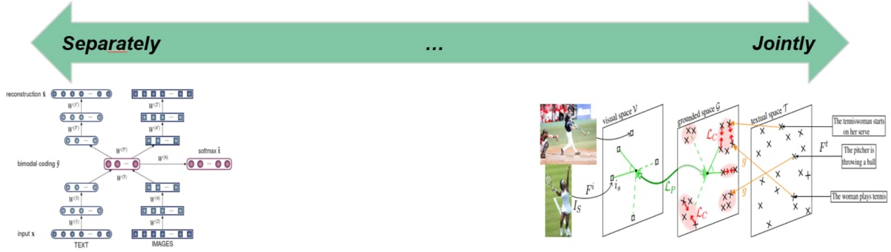

FIGURE 13.1: Left, Silberer et al., 2014: stacked autoencoders to learn higher-level embeddings from textual and visual modalities, encoded as vectors of attributes. Right, Bordes et al., 2020: textual and visual information fused in an Intermediate space denoted as “grounded space”; the “grounding objective function” is not applied directly on sentence embeddings but trained on this intermediate space, on which sentence embeddings are projected.

For example, Silberer and Lapata (2012) implement a model where a one-to-one correspondence between textual and visual space is assumed. Text and visual representations are passed to two separate unimodal encoders and both outputs are then fed to a bimodal autoencoder. On the other side, Bordes et al. (2020) propose a “text objective function” whose parameters are shared with an additional “grounded objective function”. The training of the latter takes place in what the authors called a “grounded space”, which allows to avoid the one-to-one correspondence between textual and visual space. These are just introductory examples and between these two approaches there are many shades of gray (maybe more than fifty…). These models exhibit in many instances better performance than pure language models, but they still struggle on some aspects, for example when they deal with abstract words and sentences.

Afterwards, in “Text supporting Image Models”, approaches where natural language is used as supervision for CV models are described. Intuitively these

models should be more powerful compared to models supervised solely by manually labeled data, simply because there is much more training data available. An important example for this is the CLIP model (Radford et al., 2021) with its new dataset WIT (WebImageText) comprising 400 million text-image pairs scraped from the internet.
Similar to "Text2Image" the recent successes in NLP have inspired new approaches in this field. Most importantly pre-train methods, which directly learn from raw text (e. g. GPT-n, Generative Pre-trained Transformer; Brown et al., 2020). So, CLIP stands for Contrastive Language-Image Pre-training. A transformer-like architecture is used for jointly pre-training a text encoder and an image encoder. For this the contrastive goal to correctly predict which natural language text pertains to which image inside a certain batch, is employed. Training this way turned out to be more efficient than to generate captions for images.
This leads to a flexible model, which at test time uses the learned text encoder as a "zero-shot" classifier on embeddings of the target dataset's classes. The model, for example, can perform optical character recognition, geo-location and action-recognition. Performance-wise CLIP can be competitive with task-specific supervised models, while never seeing an instance of the specific dataset before. This suggests an important step towards closing the "robustness gap", where machine learning models fail to meet the expectations set by their previous performance – especially on ImageNet test-sets – on new datasets.

Finally, "Text plus Images" discusses how text and image inputs can be incorporated into a single unifying framework in order to get closer to a general self-supervised learning model. There are two key advantages that make such a model particularly interesting. Similar to models mentioned in previous parts, devoid of human labelling, self-supervised models don't suffer from the same capacity constraints as regular supervised learning models. Nevertheless, while there have been notable advances in dealing with different modalities, it is often unclear to which extend a model structure generalizes across different modalities. Rather than potentially learning modality-specific biases, a general multipurpose framework can help increase robustness while also simplifying the learner portfolio and thereby better emulating human learning processes.
Data2vec (Baevski et al., 2022) is a new multimodal self-supervised learning model which uses a single framework for either speech, NLP or computer vision. This is in contrast to earlier models which used different algorithms for different modalities. The core idea of data2vec, developed by MetaAI, is to predict latent representations of the full input data based on a masked view of the input in a self-distillation setup using a standard transformer architecture (Baevski et al., 2022). As a result, the main improvement is in the framework, not the underlying models themselves. For example, the transformer architecture follows Vaswani et al. (2017). Transformers have several advantages over CNNs, such as encoding the relative position of features (citation needed). The

central building block of the data2vec framework is a student-teacher structure that allows the learning process to occur without supervision. To achieve this, inputs serve both as training data and as learning targets by being masked. A key issue to be aware of is model collapse, i.e the model collapsing into a constant representation. Normalization helps prevent that, as well as the domination of certain layers with high norm. The encoding, normalization and masking strategies are modality-specific. However, the learning objective remains the same across all modalities. The model is trained to predict the model representation of the original unmasked training sample. As a result of the use of self-attention in creating teacher representations, the data2vec model works with continuous and contextualized targets which are richer in information than a fixed set of targets based on local context as used in most prior work. On top of that, working with latent representations of the network itself can be seen as a simplification of many prior modality-specific models (Baevski et al., 2022). As far as the results are concerned, data2vec is effective in all three modalities. It sets new SOTA scores on computer vision, speech recognition as well as speech learning benchmarking sets.

14

title

Author:

Supervisor:

15

title

Author:

Supervisor:

16

title

Author:

Supervisor:

17

title

Author:

Supervisor:

18

title

Author:

Supervisor:

19

Epilogue

Author:

19.1 test

20

Acknowledgements

The most important contributions are from the students themselves. The success of such projects highly depends on the students. And this book is a success, so thanks a lot to all the authors! The other important role is the supervisor. Thanks to all the supervisors who participated! Special thanks to Christian Heumann[1] and Bernd Bischl[2] who enabled us to conduct the seminar in such an experimental way, supported us and gave valuable feedback for the seminar structure. Thanks a lot as well to the entire Department of Statistics[3] and the LMU Munich[4] for the infrastructure.

The authors of this work take full responsibilities for its content.

[1] https://www.misoda.statistik.uni-muenchen.de/personen/professoren/heumann/index.html

[2] https://www.statistik.uni-muenchen.de/personen/professoren/bischl/index.html

[3] https://www.statistik.uni-muenchen.de/

[4] http://www.en.uni-muenchen.de/index.html

Bibliography

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.

Bordes, P., Zablocki, E., Soulier, L., Piwowarski, B., and Gallinari, P. (2020). Incorporating visual semantics into sentence representations within a grounded space. *arXiv preprint arXiv:2002.02734*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2021). GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Silberer, C. and Lapata, M. (2012). Grounded models of semantic representation. In *Tsujii J, Henderson J, Paşca M, editors. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; 2012 Jul 12–14; Jeju Island, Korea. Stroudsburg: ACL; 2012. p. 1423-33.* ACL (Association for Computational Linguistics).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.