Multimodal Deep Learning

Contents

Preface

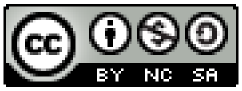

FIGURE 1: Creative Commons License

[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: Cem Akkus, Vladana Djakovic, Christopher Benjamin Marquardt

Supervisor: Dr. Matthias Aßenmacher

Natural Language Processing (NLP) has existed for about 50 years, but it is more relevant than ever. There have been several breakthroughs in this branch of machine learning that is concerned with spoken and written language. For example, learning internal representations of words was one of the greater advances of the last decade. Word embeddings (Mikolov et al. (2013), Bojanowski et al. (2016)) made it possible and allowed developers to encode words as dense vectors that capture their underlying semantic content. In this way, similar words are embedded close to each other in a lower-dimensional feature space. Another important challenge was solved by Encoder-decoder (also called sequence-to-sequence) architectures Sutskever et al. (2014), which made it possible to map input sequences to output sequences of different lengths. They are especially useful for complex tasks like machine translation, video captioning or question answering. This approach makes minimal assumptions on the sequence structure and can deal with different word orders and active, as well as passive voice.

A definitely significant state-of-the-art technique is Attention Bahdanau et al. (2014), which enables models to actively shift their focus – just like humans do. It allows following one thought at a time while suppressing information irrelevant to the task. As a consequence, it has been shown to significantly improve performance for tasks like machine translation. By giving the decoder access to directly look at the source, the bottleneck is avoided and at the same time, it provides a shortcut to faraway states and thus helps with the vanishing gradient problem. One of the most recent sequence data modeling techniques is Transformers (Vaswani et al. (2017)), which are solely based on attention and do not have to process the input data sequentially (like RNNs). Therefore, the deep learning model is better in remembering context-induced earlier in long sequences. It is the dominant paradigm in NLP currently and even makes better use of GPUs, because it can perform parallel operations. Transformer architectures like BERT (Devlin et al. (2018)), T5 (Raffel et al. (2019)) or GPT-3 (Brown et al. (2020)) are pre-trained on a large corpus and can be fine-tuned for specific language tasks. They have the capability to generate stories, poems, code and much more. With the help of the aforementioned breakthroughs, deep networks have been successful in retrieving information and finding representations of semantics in the modality text. In the next paragraphs, developments for another modality image are going to be presented.

Computer vision (CV) focuses on replicating parts of the complexity of the human visual system and enabling computers to identify and process objects in images and videos in the same way that humans do. In recent years it has become one of the main and widely applied fields of computer science. However, there are still problems that are current research topics, whose solutions depend on the research's view on the topic. One of the problems is how to optimize deep convolutional neural networks for image classification. The accuracy of classification depends on width, depth and image resolution. One way to address the degradation of training accuracy is by introducing a deep residual learning framework (He et al., 2015). On the other hand, another less common method is to scale up ConvNets, to achieve better accuracy is by scaling up image resolution. Based on this observation, there was proposed a simple yet effective compound scaling method, called EfficientNets (Tan and Le, 2019).

Another state-of-the-art trend in computer vision is learning effective visual representations without human supervision. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results, but the simple framework for contrastive learning of visual representations, which is called SimCLR, outperforms previous work (Chen et al., 2020). However, another research proposes as an alternative a simple "swapped" prediction problem where we predict the code of a view from the representation of another view. Where features are learned by Swapping Assignments between multiple Views of the same image (SwAV) (Caron et al., 2020). Further recent contrastive methods are trained by reducing the distance between representations of different augmented views of the same image ('positive pairs') and increasing the distance between representations of augmented views from different images ('negative pairs'). Bootstrap Your Own Latent (BYOL) is a new algorithm for self-supervised learning of image representatios (Grill et al., 2020).

Self-attention-based architectures, in particular, Transformers have become the model of choice in natural language processing (NLP). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention, some replacing the convolutions entirely. The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Inspired by the Transformer scaling successes in NLP, one of the experiments is applying a standard Transformer directly to the image (Dosovitskiy et al., 2020). Due to the widespread application of computer vision, these problems differ and are constantly being at the center of attention of more and more research.

With the rapid development in NLP and CV in recent years, it was just a question of time to merge both modalities to tackle multi-modal tasks. The release of DALL-E 2 just hints at what one can expect from this merge in the future. DALL-E 2 is able to create photorealistic images or even art from any given text input. So it takes the information of one modality and turns it into another modality. It needs multi-modal datasets to make this possible, which are still relatively rare. This shows the importance of available data and the ability to use it even more. Nevertheless, all modalities are in need of huge datasets to pre-train their models. It's common to pre-train a model and fine-tune it afterwards for a specific task on another dataset. For example, every state-of-the-art CV model uses a classifier pre-trained on an ImageNet based dataset. The cardinality of the datasets used for CV is immense, but the datasets used for NLP are of a completely different magnitude. BERT uses the English Wikipedia and the Bookscorpus to pre-train the model. The latter consists of almost 1 billion words and 74 million sentences. The pre-training of GPT-3 is composed of five huge corpora: CommonCrawl, Books1 and Books2, Wikipedia and WebText2. Unlike language model pre-training that can leverage tremendous natural language data, vision-language tasks require high-quality image descriptions that are hard to obtain for free. Widely used pre-training datasets for VL-PTM are Microsoft Common Objects in Context (COCO), Visual Genome (VG), Conceptual Captions (CC), Flickr30k, LAION-400M and LAION-5B, which is now the biggest openly accessible image-text dataset.

Besides the importance of pre-training data, there must also be a way to test or compare the different models. A reasonable approach is to compare the performance on specific tasks, which is called benchmarking. A nice feature of benchmarks is that they allow us to compare the models to a human baseline. Different metrics are used to compare the performance of the models. Accuracy is widely used, but there are also some others. For CV the most common benchmark datasets are ImageNet, ImageNetReaL, CIFAR-10(0), OXFORD-IIIT PET, OXFORD Flower 102, COCO and Visual Task Adaptation Benchmark (VTAB). The most common benchmarks for NLP are General Language Understanding Evaluation (GLUE), SuperGLUE, SQuAD 1.1, SQuAD 2.0, SWAG, RACE, ReCoRD, and CoNLL-2003. VTAB, GLUE and SuperGLUE also provide a public leader board. Cross-modal tasks such as Visual Question Answering (VQA), Visual Commonsense Reasoning (VCR), Natural Language Visual Reasoning (NLVR), Flickr30K, COCO and Visual Entailment are common benchmarks for VL-PTM.

2.1 title

Author:

Supervisor:

2.2 title

Author:

Supervisor:

3

Resources and Benchmarks for NLP, CV and multimodal tasks

Author: Christopher Marquardt

Supervisor: Prof. Dr. Christian Heumann

Small Intro of my chapter

- Explain that pre-training is huge part why NLP and CV models perform good
- Hint also that combination of both will be rest of the book

Intro for pretraining Ideas:

- like an athlete.
- Need some base fitness (=pre-training)
- Same like in reality pre-training differs between models.

3.0.1 Pre-training

3.0.1.1 Resources for pre-training

- how does pre-training look for NLP, CV, MML
- State and explain 3-4 of the most used resources (maybe add more)
- (How much effort to clean pre-training data)
- provide resources to find more (papers with code, ...)
- availability and size of pre-training for different modalities
  - NLP > CV > MML (MML pretty new compared to others)
  - Most of them not public (not good; Example JFT-300M)
  - What role does size play (logarithmic)

- How has pre-training changed or has it even changed in modalities
  - CV: still all train on ImageNet ("Are we done with ImageNet") and poor performance of ObjectNet
  - use of noisy data; ?quantity > quality?

3.0.1.2 Use of resources

- How pre-training is used in different modalities
  - supervised
  - self-supervised
- State and explain 2 or 3 main used pre-training tasks
  - masked approaches
  - ...

3.0.2 Fine-tuning

- why fine-tuning is important
- where and how to fine-tune
- ...

3.0.3 Benchmarks for modalities

- Importance of benchmarks
  - also need for new ones (like Psych: Flynn Effect)
    - do models get better or is it possible that pre-training contains already benchmarks (too much crawl; NLP)
  - Hint that models pre-train on different resources but perform on same benchmarks (good or bad)
- different tasks for benchmarks
  - state most important ones also give infor to find ohters (example: papers with code)
  - Hint that it's most of the time reduction to classification tasks (like is this next sentences)
  - Semantic of produced sentences often not nice

Outro: Multimodal architectures Chapter

4

Chapter 1

Authors: Author 1, Author 2

Supervisor: Supervisor

4.1 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)

4.2 Using Figures

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

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

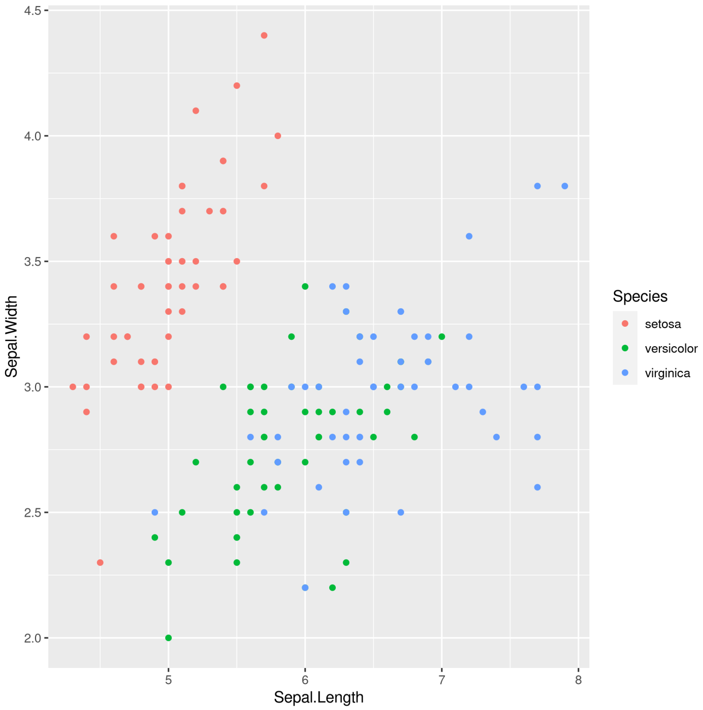

FIGURE 4.1: This is the caption of the figure!

4.3 Using Tex

HTML rendering uses MathJax while pdf rendering uses LaTeX:

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

4.4 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

5

title

Author:

Supervisor:

6

title

Author:

Supervisor:

7

title

Author:

Supervisor:

8

title

Author:

Supervisor:

9

title

Author:

Supervisor:

10

title

Author:

Supervisor:

11

Chapter 2 Multimodal architectures

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

Supervisor: Christian Heumann

11.1 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.

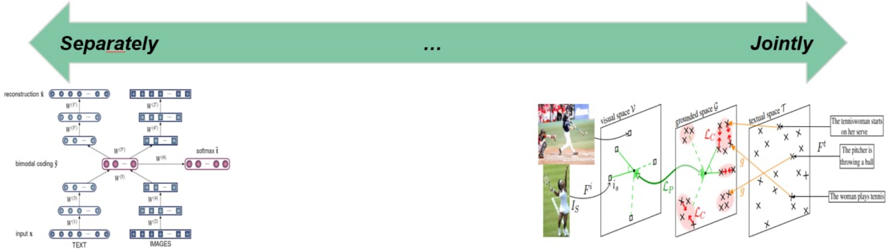

FIGURE 11.1: Left, Silberer et al., 2014: stacked autoencoders to learn higher-level embeddings from textual and visual modalities, encoded as vectors of attributes. Right, Bordes et al., 2020: textual and visual information fused in an Intermediate space denoted as “grounded space”; the “grounding objective 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.

12

Further Topics

Authors: Marco Moldovan, Rickmer Schulte, Philipp Koch

Supervisor: Rasmus Hvingelby

So far we have learned about multimodal models for text and 2D images. Text and images can be seen as merely snapshots of the sensory stimulus that we humans perceive constantly. If we view the research field of multimodal deep learning as a means to approach human-level capabilities of perceiving and processing real-world signals then we have to consider lots of other modalities in a trainable model other than textual representation of language or static images. Besides introducing further modalities that are frequently encountered in muli-modal deep learning, the following chapter will also aim to bridge the gap between the two fundamental sources of data, namely structured and unstructured data. Investigating modeling approaches from both classical statistics and more recent deep learning we will examine the strengths and weaknesses of those and will discover that a combination of both may be a promising path for future research. Going from multi modalities to multi task, the last section will then broaden our view of multi-modal deep learning by examining multi purpose modals. Discussing cutting-edge research topics such as the newly developed Pathways, we will discuss current achievements and limitations of the new modeling and hardware approaches that might lead our way towards the ultimate goal of AGI in multi-modal deep learning.

13

Further Modalities

Author: Marco Moldovan

Supervisor: Rasmus Hvingelby

13.1 Intro

In this chapter we will build up a taxonomy of different perceivable and interpretable types of signals that we as humans use to navigate the world and we will see how today's state-of-the-art multimodal models are built and trained in order to process more and more modalities simultaneously in order to build more and more complete representations of world through available data. We will build up our taxonomy starting from the two most well-understood modalities - namely text and 2D images - and introduce models that learn relationships between increasingly many modalities at the same time and to map them to a cross-modal representation space in which we can apply distance functions to points in order to represent semantic relatedness between datapoint from these different modalities. Given such a learned cross-modal representation space we will look at some of the most important multimodal downstream tasks and applications.

Towards the end of the chapter we will take a closer look at the two main types of model architectures and training paradigms: bi-encoders and "true" multimodal cross-encoders. The first kind of model can be seen as an ensemble of unimodal expert models that map into the same representation space while using some form of metric learning to relate representations of different modalities to one another. True multimodal models are essentially agnostic to their input (as long as it is preprocessed and featurized appropriately). We currently see the second kind of architecture as the more promising one for the case of approximating human-level perception. An example of a modality agnostic multimodal model is the Perceiver of which we will introduce a newer, even more efficient variant.

Up until recently each modality required their own specific self-supervised training paradigm: for text a common approach would be MLM while the same training paradigm wasn't as effective for images or video. data2vec introduces a modality-agnostic SSL masked prediction setup which requires careful preprocessing but does not care about the source of the input. We see a model that marries a modality-agnostic model like Perceiver with a modality agnostic training paradigm like data2vec as a very promising path forward. Topic 11 will build on this idea of modality-agnostic models by introducing Google's Pathways: a concept for multimodal, multi-task, sparse world models.

14

Strucutered + Unstrucutered Data

Author: Rickmer Schulte

Supervisor: Daniel Schalk

14.1 Intro

While the previous chapter has extended the range of modalities considered in multi-modal deep learning beyond image and text data, the focus remained on other sorts of unstructured data. This has neglected the broad class of structured data, which has been the basis for research in pre deep learning eras and which has given rise to many fundamental modeling approaches in statistics and classical machine learning. Hence, the following chapter will aim to give an overview of both data sources and outline the respective ways these have been used for modeling purposes as well as more recent attempts to model them jointly.

Generally, structured and unstructured data substantially differ in certain aspects such as dimensionality and interpretability which have led to various modeling approaches that are particularly designed for the special characteristics of the data types, respectively. As shown in previous chapters, deep learning models such as neural networks are known to work well on unstructured data due to their ability to extract latent representation and to learn complex dependencies from unstructured data sources to achieve state-of-the art performance on many classification and prediction tasks. By contrast, classical statistical models are mostly applied to tabular data due the advantage of interpretability inherent to these models, which is commonly of great interest in many research fields. However, as more and more data has become available to researchers today, they often do not only have one sort of data modality at hand but both structured and unstructured data at the same time. Discarding one or the other data modality makes it likely to miss out on valuable insights and potential performance improvements.

Therefore, the following chapter will mainly examine different concepts to model both data types jointly and discuss under which circumstances one or the other is better suited. Methods such as feature engineering to integrate unstructured data via expert knowledge into the classical model framework as well as different fusion strategies to integrate both types of modalities into common deep learning architectures are analyzed and evaluated. Especially the latter will be explored in detail by referring to numerous examples from health care, biology and finance. Finally, recently proposed methods will be discussed, which bring together classical statistical models for structured data and neural networks for unstructured data in order to yield interpretable deep learning models that combine the best of both worlds.

15

Multi-purpose Models

Author: Philipp Koch

Supervisor: Rasmus Hvingelby

15.1 Intro

After we describe further modalities in the previous sections, we will look at truly multipurpose models. Multitask multimodal models have already been proposed, like UniT, which extends the transformer architecture to deal with different modalities and tasks. However, previous multitask multimodal models remain limited in different aspects, which we will describe and discuss further. To become genuinely multipurpose, however, a model must be able to solve different tasks without fine-tuning and must be capable of dealing with different modalities. Thus, it must be able to transfer knowledge in-between tasks but must also be able to allocate capabilities for different modalities.

The recently introduced deep learning architecture Pathways is designed to be multipurpose. Pathways builds on newly designed hardware and software dedicated to addressing the challenges of contemporary deep learning models, which are ever-growing, where GPT-3 might be the most prominent example. We will discuss previous drawbacks and describe how Pathways aims to solve these issues. Besides the hardware aspect, Pathways provides a large neural network constructed as a directed acyclic graph (DAG). The input is passed through the network on different paths. Each node of the network is itself a neural network aimed at solving a specific aspect of a task. Using these different neural networks inside the model allows the model to be multitask and transfer knowledge in-between tasks. Another important aspect of this architecture is the obtained sparsity. When computed, just necessary nodes are computed, resulting in higher overall performance.

Furthermore, the model is intended to absorb different modalities as input, where no implementation has been found. Multimodality is further used hypothetically in the initial blog post. However, the similar model PathNet also achieves multimodality. The only model based on Pathways is the language model (PaLM), which is multilingual and capable of understanding code and solving mathematical tasks. However, the multimodality here remains questionable. Future Pathways-based models might provide more insight if the claim to step further toward artificial general intelligence (AGI) of the authors of Pathways and PathNet is true or not. Eventually, we will discuss the impact of the new Pathways multipurpose model since it might have a large impact on deep learning in the upcoming future. Broader applicable models will become feasible yet also centralize the usage, thus reducing accessibility and subsequently research on these models.

16

title

Author:

Supervisor:

17

Epilogue

Author:

17.1 test

18

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*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information.

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.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

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*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

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.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

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.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

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).

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

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.