

Challenges and Perspectives in Deep Generative Modeling

Vincent Fortuin^{*1}, Yingzhen Li^{*2}, Kevin Murphy^{*3},
Stephan Mandt^{*4}, and Laura Manduchi^{†5}

- 1 University of Cambridge, GB. vbf21@cam.ac.uk
- 2 Imperial College London, GB. yingzhen.li@imperial.ac.uk
- 3 Google Research – Mountain View, US. kpmurphy@google.com
- 4 University of California – Irvine, US. mandt@uci.edu
- 5 ETH Zürich, CH. laura.manduchi@inf.ethz.ch

Abstract

Deep generative models, such as variational autoencoders, generative adversarial networks, normalizing flows, and diffusion probabilistic models, have attracted a lot of recent interest. However, we believe that several *challenges* hinder their more widespread adoption: (C1) the difficulty of objectively evaluating the generated data; (C2) challenges in designing scalable architectures for fast likelihood evaluation or sampling; and (C3) challenges related to finding reproducible, interpretable, and semantically meaningful latent representations. In this Dagstuhl Seminar, we have discussed these open problems in the context of real-world *applications* of deep generative models, including (A1) generative modeling of scientific data, (A2) neural data compression, and (A3) out-of-distribution detection. By discussing challenges C1–C3 in concrete contexts A1–A3, we have worked towards identifying commonly occurring problems and ways towards overcoming them. We thus foresee many future research collaborations to arise from this seminar and for the discussed ideas to form the foundation for fruitful avenues of future research. We proceed in this report by summarizing the main results of the seminar and then giving an overview of the different contributed talks and working group discussions.

Seminar February 12–17, 2023 – <https://www.dagstuhl.de/23072>

2012 ACM Subject Classification Computing methodologies → Unsupervised learning; Computing methodologies → Kernel methods; Computing methodologies → Learning in probabilistic graphical models

Keywords and phrases deep generative models, representation learning, generative modeling, neural data compression, out-of-distribution detection

Digital Object Identifier 10.4230/DagRep.13.2.47

1 Executive Summary

Vincent Fortuin (University of Cambridge, UK)

Yingzhen Li (Imperial College London, UK)

Kevin Murphy (Google Research – Mountain View, US)

Stephan Mandt (University of California – Irvine, US)

License © Creative Commons BY 4.0 International license
© Vincent Fortuin, Yingzhen Li, Kevin Murphy, and Stephan Mandt

Premise

Since the inception of variational autoencoders, generative adversarial networks, normalizing flows, and diffusion models, the field of deep generative modeling has grown rapidly and consistently over the years. Especially in recent years, this has led to great advances in

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Challenges and Perspectives in Deep Generative Modeling, *Dagstuhl Reports*, Vol. 13, Issue 2, pp. 47–70

Editors: Vincent Fortuin, Yingzhen Li, Kevin Murphy, Stephan Mandt, and Laura Manduchi



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

generating images, speech and text, as well as great promises in generating structured data such as 3D objects, videos, and molecules. However, we believe that current research has not sufficiently addressed several fundamental challenges related to evaluating and scaling these models, as well as interpreting their latent structure. These challenges have different manifestations in different applications. For example, while a variational autoencoder’s sensitivity to changing data distributions can induce long code lengths and poor image reconstructions in neural compression, the same feature can be a positive attribute in detecting anomalies.

We believe that it is most beneficial to understand the challenges of deep generative models in their practical contexts. For this reason, we have invited a combination of researchers working on foundations of generative models and researchers working on specialized applications to this Dagstuhl Seminar. Thus, by integrating different communities, we have made a step towards identifying generalizable solutions across domains that spur innovation and new research.

As the main challenges of current deep generative modeling approaches we have identified the evaluation of generative models, performing scalable inference in such models, and improving the interpretability and robustness of the models’ learned latent representations.

As example applications, we have considered three *application areas* that draw on generative modeling and that show various manifestations of the aforementioned *challenges*. Concretely, we consider applications in modeling scientific data, neural data compression, and out-of-distribution detection.

Structure of the seminar

We have created an open and inclusive atmosphere where participants from different communities could mingle and exchange ideas, leaving enough room for serendipitous encounters and ad-hoc discussions. We have catalyzed this process by inviting the participants to give short talks on either models (for the researchers) or problems (for the practitioners) as a basis for subsequent discussions. We then had panel discussions and round-tables regarding different topics that the participants could self-assign to, in order to match their common interests.

To promote interactions among researchers especially between those who may have not known each other, we have randomly paired researchers and practitioners into pairs and small groups and assigned them small tasks, such as coming up with a short abstract that would combine their interests. These types of activities have ultimately planted the seeds for different future collaborations and fostered a sense of togetherness among the participants.

Main observations from the talks

The content of the talks is covered in more detail in the next section, but we want to take the opportunity here to highlight recurring patterns and topics that emerged.

One main observation was that while large generative models, such as diffusion models or large language models, yield impressive performance and can solve many tasks that we would naïvely not have expected them to solve well (e.g., diffusion models sorting lists or solving sudokus and large language models performing logical reasoning), we lack a proper

theoretical understanding of these models and can thus not guarantee their safety or reliability. This makes it particularly dangerous to use these models in critical applications, such as healthcare.

Moreover, many domains have specific requirements that are well-known to practitioners, but often ignored by machine learning researchers, e.g., non-iid data, safety constraints, prior knowledge, interpretability, or causal assumptions. While there are sub-fields of machine learning research studying these problems, most off-the-shelf methods do not readily provide solutions.

Finally, generative modeling holds great promise for areas such as neural compression or anomaly/out-of-distribution detection, but the practical improvements achieved by generative approaches in these domains remain limited. We will need more targeted collaborations between experts in generative modeling and these problem settings to make tangible real-world progress, some of which will have hopefully been sparked by this seminar.

Main takeaways from the working groups

Our working group sessions self-assembled spontaneously around key topics of interest that had emerged from the talks and informal discussions during the breaks. They focused on prior knowledge, continual learning, and anomaly detection.

Firstly, when it comes to domain knowledge, one working group tried to develop a categorization of different types and came up with physical constraints, symmetries, logic, ontologies, and factual knowledge. All of these require different approaches to incorporate them into generative models, so the developers of the model should be cognizant of the type of domain knowledge the practitioners might have. Moreover, eliciting the prior knowledge from the experts can be hard and cumbersome, and an elicitation strategy should be designed together with the model itself.

Secondly, continual learning is well-studied in the supervised setting, but less so in the unsupervised one. However, in the age of large generative models that are very expensive to train, continually expanding their generative abilities without having to retrain them from scratch becomes paramount. Since no explicit supervised objective function is available to measure the learning progress or potential forgetting, new solutions need to be developed to efficiently learn continually without catastrophic forgetting in the generative context.

Lastly, anomaly detection is a hard problem that has been studied in the statistical literature for decades, but novel powerful generative models harbor the promise of estimating quantities such as the compressibility or Kolmogorov complexity of data points, which might be used to more effectively detect outliers, out-of-distribution examples, and anomalous inputs.

2 Table of Contents

Executive Summary

| | |
|--|----|
| <i>Vincent Fortuin, Yingzhen Li, Kevin Murphy, and Stephan Mandt</i> | 47 |
|--|----|

Overview of Talks

| | |
|--|----|
| Large Language Models vs. Large AI Models <i>Gerard de Melo</i> | 52 |
| Image membership in generative models <i>Sina Däubener</i> | 53 |
| Disentangling Style and Content for Neural Topic Models <i>Sophie Fellenz</i> | 53 |
| Modeling mixed-tailed distributions with Normalizing Flow and convergence of the ELBO of VAEs to a sum of three entropies <i>Asja Fischer</i> | 54 |
| Gaussian Process Variational Autoencoders <i>Vincent Fortuin</i> | 55 |
| Active search in structured spaces with domain-specific similarities <i>Thomas Gärtner</i> | 55 |
| Deep Generative Models in Healthcare <i>Julia Vogt</i> | 56 |
| Do we care about non-iid data for generative models? <i>Matthias Kirchler</i> | 56 |
| CMSSG: Heterogeneous image data integration with Causal Multi-Source StyleGAN <i>Christoph Lippert</i> | 56 |
| Self-Supervised Learning beyond Vision and Language <i>Maja Rudolph</i> | 58 |
| Towards Runtime-Efficient Neural Compression <i>Stephan Mandt</i> | 58 |
| Informed Representation Learning with Deep Generative Models <i>Laura Manduchi</i> | 59 |
| Towards Anytime Computation in Deep Architectures <i>Eric Nalisnick</i> | 61 |
| The Future (R)evolution of Generative AI <i>Björn Ommer</i> | 61 |
| Where to Diffuse, how to diffuse, and how to get back? <i>Rajesh Ranganath</i> | 62 |
| Universal Critics <i>Lucas Theis</i> | 62 |
| Getting the most **out** of your representations <i>Karen Ullrich</i> | 63 |
| Interventional causal representation learning with deep generative models <i>Yixin Wang</i> | 63 |

| | |
|--|----|
| Assaying Out-Of-Distribution Generalization in Transfer Learning <i>Florian Wenzel</i> | 64 |
| Trading Information between Latents in Hierarchical Variational Autoencoders <i>Robert Bamler</i> | 65 |
| Challenges in Generative Language Modeling <i>Alexander Rush</i> | 66 |
| Fun with Foundation Models and Amortized Inference <i>Frank Wood</i> | 66 |
| Languages for the Next 700 Application Domains in AI <i>Jan-Willem van de Meent</i> | 67 |
| Working groups | |
| Continual learning of deep generative models <i>Sophie Fellenz, Sina Däubener, Gerard de Melo, Florian Wenzel, and Frank Wood</i> | 67 |
| Priors in deep generative modeling <i>Vincent Fortuin, Thomas Gärtner, Matthias Kirchler, and Eric Nalisnick</i> | 68 |
| The role of domain knowledge in deep generative models <i>Vincent Fortuin, Thomas Gärtner, Matthias Kirchler, Christoph Lippert, Laura Manduchi, Guy Van den Broeck, Julia Vogt, and Florian Wenzel</i> | 69 |
| Anomaly detection using Kolmogorov complexities <i>Marius Kloft, Asja Fischer, and Lucas Theis</i> | 69 |
| Participants | 70 |

3 Overview of Talks

3.1 Large Language Models vs. Large AI Models

Gerard de Melo (*Hasso-Plattner-Institut, Universität Potsdam, DE*)

License  Creative Commons BY 4.0 International license
 Gerard de Melo

Language models and other generative models of symbolic sequences have a long history that can be traced back to early studies on prediction probabilities for written language such as those by Shannon [6]. Later on, in the 1980s, their statistics started becoming crucial components of systems for machine translation, speech recognition, and optical character recognition. In the 2000s, the widespread availability of Web-scale word n-gram statistics opened up many new opportunities for language model-driven applications [3].

To generate outputs that resemble human-written text remarkably well at a superficial level, it suffices to sample from very simple n-gram language models. With the advent of neural network-driven language models [1], the generalization abilities improved further. Finally, the series of milestone successes of large language models, most notably the powerful GPT family of neural models [4], has led to generative models of language with unprecedented abilities to generalize to new tasks simply by following instructions.

Along with these improved prediction capabilities, I argue that another notable paradigm shift has emerged. Early studies had already shown that language models are not just mere models of linguistic well-formedness, but rather valuable sources of knowledge [7, 8]. With the powerful capabilities of the GPT models, people started to expect them to serve as universal engines for question answering and broader AI tasks.

While language models are normally supposed to produce *plausible* text, current models are increasingly expected to satisfy a number of *additional desiderata*. For instance, there is a need for models that provide only statements that are deemed factually accurate and trustworthy [2]. There are widespread calls for such models to avoid toxicity and bias [5]. With their deployment in commercial search engines and other mainstream applications, current models are expected to avoid outputs that may lead to harmful effects, for instance by refraining from responding in ways that could pose a risk for the mental health of human interlocutors and by refusing to carry out tasks related to illegal activities.

Thus, large language models are no longer just models of language but general-purpose AI models, leading to an urgent need for us to develop improved generative modeling techniques with substantially better constraint satisfaction and uncertainty estimation.

References

- 1 Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* vol. 3 (2003), pp. 11-1155. 2003.
- 2 Rajarshi Bhowmik, Gerard de Melo. Be Concise and Precise: Synthesizing Open-Domain Entity Descriptions from Facts. *Proceedings of The Web Conference 2019*. ACM, 2019.
- 3 Thorsten Brants, Alex Franz. *Web 1T 5-gram Version 1*. LDC2006T13. Linguistic Data Consortium, Philadelphia, 2006.
- 4 OpenAI. GPT-4 Technical Report. *arXiv* 2303.08774. 2023.
- 5 Lena Schwertmann, Manoj Prabhakar Kannan Ravi, Gerard de Melo. Model-Agnostic Bias Measurement in Link Prediction. *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, 2023.
- 6 Claude E. Shannon. Prediction and entropy of printed English. *The Bell System Technical Journal* vol. 30, no. 1, pp. 50–64. IEEE, 1951.

- 7 Niket Tandon, Gerard de Melo. Information Extraction from Web-Scale N-Gram Data. In: *Proceedings of the Web N-gram Workshop at SIGIR 2010*. ACM, 2010.
- 8 Niket Tandon, Gerard de Melo, Gerhard Weikum. Deriving a Web-Scale Common Sense Fact Database. In: *Proceedings of AAAI 2011*. AAAI, 2011.

3.2 Image membership in generative models

Sina Däubener (Ruhr-Universität Bochum, DE)

License © Creative Commons BY 4.0 International license
© Sina Däubener

Joint work of Sina Däubener, Asja Fischer, Mike Laszkiewicz, Denis Lukovnikov, Jonas Ricker, Simon Damm, Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Dorothea Kolossa, Thorsten Holz

Images from current state-of-the-art generative models have unarguably led to impressive results. However, this comes with potential threats such as the spread of misinformation through generated images or unauthorized style imitations of certain artists.

To tackle the first problem, I'll first present works from our group and collaborators, which do successful deep fake detection in the frequency domain of images. In the second part, I'll talk about current work in progress where we try to (invisibly) watermark images, such that a model fine tuned on these images picks up on the watermark. This would allow us to certify that works from a certain artist have in fact been used.

3.3 Disentangling Style and Content for Neural Topic Models

Sophie Fellenz (RPTU – Kaiserslautern, DE)

License © Creative Commons BY 4.0 International license
© Sophie Fellenz

Joint work of Sophie Fellenz, Mayank Kumar Nagda

Neural topic models are used to quickly get an overview of the central themes in large text corpora. They are typically trained only on content words, whereas other words related to syntax or style of the text are removed in preprocessing. Rather than relying on manually curated and static stop word lists, we propose a data-driven way of dynamically identifying the style component in text data. The idea is to differentiate between long-range and short-range dependencies in text data. Following previous work in linguistics and cognitive science we posit that content words tend to have long-range dependencies whereas style or syntax words have short-range dependencies. Topic models are good at learning topics by disregarding sequential information, exclusively focusing on long-range dependencies. Language models however process text sequentially to predict the next word given the immediate short-range context. In this talk I present a method for combining both types of models to automatically distinguish syntactic and semantic words. Instead of simply removing the syntactic words, we can also group them and use them for the text analysis alongside the semantic words. Results show that this data-driven way of separating the words leads to higher topic quality and better feature selection on a range of datasets. This may lead to better ways of disentangling content and style in text data in the future, aiding controlled text generation for longer texts and overcoming the current bottleneck in text generation which is the size of input prompt. It was discussed how the topics in neural topic models can be seen as experts and the topic distribution of each document as a product of experts. This view point could help

to develop topic models with a flexible number of topics where topics could be added as new experts. An interesting direction might be to look at hierarchical topic models as a collection of experts where more general experts are located at the top of the hierarchy and combinations of experts or more specific experts would be at the bottom of the hierarchy. Furthermore, topic models could be integrated with foundation models for language. To do this, the tokenization and input format for topic models need to be unified and a general mechanism for conditioning on and extracting topics during text generation needs to be developed.

3.4 Modeling mixed-tailed distributions with Normalizing Flow and convergence of the ELBO of VAEs to a sum of three entropies

Asja Fischer (Ruhr-Universität Bochum, DE)

License  Creative Commons BY 4.0 International license
© Asja Fischer

Main reference Jörg Lücke, Dennis Forster, Zhenwen Dai: “The Evidence Lower Bound of Variational Autoencoders Converges to a Sum of Three Entropies”, CoRR, Vol. abs/2010.14860, 2020.

URL <https://arxiv.org/abs/2010.14860>

Main reference Mike Laszkiewicz, Johannes Lederer, Asja Fischer: “Marginal Tail-Adaptive Normalizing Flows”, in Proc. of the International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, Proceedings of Machine Learning Research, Vol. 162, pp. 12020–12048, PMLR, 2022.

URL <https://proceedings.mlr.press/v162/laszlkiewicz22a.html>

This talk highlights two recent theoretical results about deep generative models.

The first part is based on the work of Laszkiewicz et al. (2022) and introduces an approach for normalizing flows that allows to model distributions with heavy- as well as light-tailed marginals. We prove that the marginal tailedness of an autoregressive flow can be controlled via the tailedness of the marginals of its base distribution (i.e. the distribution of the hidden variables). This theoretical insight leads us to a novel type of flows that are based on a three-step procedure: first, estimating the marginal tail indices, second, accordingly defining a set of heavy-tailed and a set of light-tailed base distribution, and third, training a normalizing flow with data-driven linear layers.

The second part is based on recent work Damm et al. (2023). Here we show that for standard (i.e., Gaussian) VAEs the ELBO converges to a value given by the sum of three entropies: the (negative) entropy of the prior distribution, the expected (negative) entropy of the observable distribution, and the average entropy of the variational distributions (the latter is already part of the ELBO). The result implies that the ELBO can for standard VAEs often be computed in closed-form at stationary points while the original ELBO requires numerical approximations of integrals.

Both works serve as illustrative examples of the importance of improving our theoretical understanding of deep generative models to gain robust, exact, and efficient generative models.

3.5 Gaussian Process Variational Autoencoders

Vincent Fortuin (*University of Cambridge, GB*)

License © Creative Commons BY 4.0 International license
© Vincent Fortuin

Joint work of Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, Stephan Mandt
Main reference Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, Stephan Mandt: “GP-VAE: Deep Probabilistic Time Series Imputation”, in Proc. of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, Vol. 108, pp. 1651–1661, PMLR, 2020.

URL <https://proceedings.mlr.press/v108/fortuin20a.html>

Variational autoencoders (VAEs) are performant deep generative models, based on principled Bayesian inference techniques. However, in practice, people often use them with isotropic Gaussian priors over the latents, which makes all the different data points independent from each other. This often does not match our true prior beliefs, especially when working with time series data. In this talk, I gave a brief overview of some recent approaches using Gaussian processes (GPs) as priors in VAEs, highlighting their tradeoffs and historical development. During the discussion, we discussed some newer follow-up work that uses Kalman filters for the GP inference in the latent space. We also discussed the choice of kernel and what the computational tradeoffs are between a full variational GP and a variational Gauss-Markov process.

3.6 Active search in structured spaces with domain-specific similarities

Thomas Gärtner (*Technische Universität Wien, AT*)

License © Creative Commons BY 4.0 International license
© Thomas Gärtner

Generating structured data has many important real-world applications such as generating molecular graphs for (de novo) drug discovery or genome sequences of bacteriophages to combat antibiotic resistant bacteria. The design space of potentially interesting structures in such applications is typically huge, it has, for instance, been estimated that there are more than 10^{30} different phages and even more small, drug-like molecules. This is in stark contrast to the amount of structures known to have desired properties in these domains which is often less than one hundred; in a lead discovery task for an antagonist of a particular integrin thought to play a key role in idiopathic pulmonary fibrosis 24 compounds were known to bind well from previous biological assays. For such applications, we propose an efficient active learning algorithm for generative models with domain-specific similarity measures. Similarity measures defined by domain experts are often not positive semi-definite and thus cannot be utilised by Hilbert-space kernel methods. They instead require more general Krein-space kernel methods which admit efficient learning by adapting Nyström approximations. Our active learning algorithm adapts the distribution of generated structures using a learned conditional exponential family model and leads to diverse sets of novel structures.

3.7 Deep Generative Models in Healthcare

Julia Vogt (ETH Zürich, CH)

License © Creative Commons BY 4.0 International license
© Julia Vogt

In recent years, enormous progress has been made to gather as much information as possible about an individual patient. The continuous adoption and integration of electronic medical records, linkage of data sources, and the advent of new diagnostic and digital monitoring technologies have led to an overwhelming amount of heterogeneous and multimodal clinical data. The ultimate aim is to utilize all this vast information for a medical treatment tailored to an individual patient's needs. To achieve this goal, we develop new generative machine learning techniques capable of dealing with the challenges arising in the medical application domain. The methods we develop cover for example multimodal data integration, transparent model development, or probabilistic clustering models.

3.8 Do we care about non-iid data for generative models?

Matthias Kirchler (Hasso-Plattner-Institut, Universität Potsdam, DE)

License © Creative Commons BY 4.0 International license
© Matthias Kirchler

Joint work of Matthias Kirchler, Christoph Lippert, Marius Kloft
Main reference Matthias Kirchler, Christoph Lippert, Marius Kloft: “Training Normalizing Flows from Dependent Data”, CoRR, Vol. abs/2209.14933, 2022.
URL <https://doi.org/10.48550/arXiv.2209.14933>

Current learning algorithms for generative models generally assume that data points are sampled independently, an assumption that is frequently violated in practice. We propose a likelihood objective of normalizing flows incorporating dependencies between the data points, for which we derive a flexible and efficient learning algorithm suitable for different dependency structures. Respecting dependencies between observations can improve empirical results on synthetic and real-world data, leading to higher statistical power in a downstream application to genome-wide association studies. So far, we have focused on normalizing flows due to their explicit likelihood modeling, but we can extend similar modeling approaches to other generative models. We have also assumed that the dependency structure is at least partially known in advance – we work on relaxing that assumption and learning low-rank approximations of dependency structure from the data.

3.9 CMSSG: Heterogeneous image data integration with Causal Multi-Source StyleGAN

Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE)

License © Creative Commons BY 4.0 International license
© Christoph Lippert

Joint work of Wei-Cheng Lai, Matthias Kirchler, Hadya Yassin, Jana Fehr, Alexander Rakowski, Hampus Olsson, Ludger Starke, Jason M. Millward, Sonia Waiczies, Christoph Lippert

Introduction: Generative Adversarial Networks (GANs) have emerged as powerful tools for generating realistic images, with conditional and causal models enabling fine-grained control over latent factors. In the medical domain, data scarcity and the need to integrate information

from diverse sources present challenges for existing generative models, often resulting in low-quality images ill-suited for medical applications. To address this issue, we propose the Causal Multi-Source StyleGAN (CMSSG), an algorithm that leverages prior knowledge over the data distribution in the form of causal graphs over image covariates and conditioning to integrate heterogeneous data sources with differing underlying distributions. CMSSG learns from multiple data sources with divergent causal structures in parallel, effectively synthesizing the learned distributions for data generation. We present a proof-of-concept experiment demonstrating CMSSG’s ability to generate hand-written digit images with varying morphological features. Additionally, we apply CMSSG to generate brain MR images with heterogeneous characteristics from the UK Biobank and the Alzheimer’s Disease Neuroimaging Initiative (ADNI) datasets, illustrating its capacity to capture brain anatomical variations. Our proposed algorithm offers a promising direction for unbiased data generation from disparate sources.

Methodology: Our approach consists of two components: the causal component and the multi-source component. The causal component learns causal relationships between clinical and demographic characteristics and brain MRI features, facilitating the synthesis of images with specific attributes. The multi-source component enables the generation of synthetic data from multiple datasets with distinct causal models, fostering the creation of more diverse data with various characteristics from different distributions.

CMSSG learns from multiple data sources with divergent causal structures in parallel, effectively synthesizing the learned distributions for data generation. To the best of our knowledge, our work is the first to address multi-source heterogeneity in GANs within a principled causal framework.

Experiments and Results: We validate our CMSSG method by generating hand-written digits with distinct morphological features. We then apply CMSSG to generate brain MRIs with specific clinical and demographic characteristics. We train our model to learn causal relationships from two datasets in parallel: the UK Biobank dataset, focusing on the relationships between demographic characteristics and MRIs, and the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort, examining the relationships between clinical dementia features and MRIs. Using CMSSG, we generate synthetic brain MRIs with controlled age, sex, brain volumes, and cognitive function (normal or impaired).

Our results demonstrate that CMSSG can synthesize high-resolution brain MRIs while realistically manipulating causal structures within the images. Although the Frechet Inception Distance (FID) score from CMSSG does not outperform CausalStyleGANs with a single causal model, it provides a new opportunity to manipulate multi-source causal covariates.

Conclusion: In conclusion, the Causal Multi-Source StyleGAN (CMSSG) represents a novel approach to address the challenges of data scarcity and biased datasets in medical image generation. By leveraging causal graphs and conditioning, CMSSG integrates heterogeneous data sources with differing underlying distributions to generate high-quality, diverse medical images. Our experiments demonstrate the efficacy of CMSSG in generating hand-written digit images and brain MRIs with specific clinical and demographic characteristics.

Future work will involve collaboration with medical experts for comprehensive quality assessment and exploration of potential applications of synthetic medical images. Additionally, further experiments should be conducted to improve the quality of synthetic images from joint causal covariates and design appropriate metrics for evaluating the causality of GANs in respect to anatomical factors. This will ensure that the model learns the correct anatomical pattern with the aging process.

3.10 Self-Supervised Learning beyond Vision and Language

Maja Rudolph (Cornell University – Ithaca, US)

License © Creative Commons BY 4.0 International license
© Maja Rudolph

Self-supervised learning has emerged as a powerful paradigm for machine learning, especially for drawing insights from unlabeled data. The key idea is to introduce auxiliary prediction tasks and to train a deep model to solve these auxiliary tasks. If the tasks are designed well, the trained model will be useful for a number of purposes, such as anomaly detection, feature extraction, and forecasting. Unfortunately, most successful approaches for SSL rely on domain-specific inductive biases and are, therefore, limited to individual use cases. In this talk, I present advanced self-supervised learning losses that facilitate domain-general self-supervised learning beyond images and text. Exponential family embeddings, for example, generalize word embeddings to provide insight into a wide range of applications. They are a useful tool for studying zebrafish brains in neuroscience, studying shopping behavior in economics, or studying language evolution in computational social science. Similarly, neural transformation learning (NTL) is a new general-purpose tool for self-supervised anomaly detection. While related methods in computer vision typically require image transformations such as rotations, blurring, or flipping, NTL automatically learns the best transformations from the data and generalizes self-supervised AD to almost any data type.

3.11 Towards Runtime-Efficient Neural Compression

Stephan Mandt (University of California – Irvine, US)

License © Creative Commons BY 4.0 International license
© Stephan Mandt

Joint work of Stephan Mandt, Yibo Yang, Robert Bamler

Main reference Yibo Yang, Robert Bamler, Stephan Mandt: “Improving Inference for Neural Image Compression”, in Proc. of the Advances in Neural Information Processing Systems, Vol. 33, pp. 573–584, Curran Associates, Inc., 2020.

URL https://proceedings.neurips.cc/paper_files/paper/2020/file/066f182b787111ed4cb65ed437f0855b-Paper.pdf

Neural image and video compression models have proven superior performance in rate-distortion and rate-perception tradeoffs compared to their classical counterparts. However, while most research still focuses on improving rate-distortion tradeoffs, neural compression models are currently much too resource-inefficient to deploy in real-world environments. This talk seeks to review strategies to maintain the strong performance of neural compression methods while aiming to reduce their runtime efficiency by 1-2 orders of magnitude. To this end, we propose three modeling and algorithmic improvements: (1) introducing lightweight decoders, (2) improving encoding at training time using semi-amortized variational inference, and (3) establishing probabilistic circuits as new models for efficient entropy coding in lossy compression.

Background: The internet and the world’s IT systems could not exist in their current form without data compression. With video streaming dominating consumer internet traffic, every percent of gained performance improvement will have a large economic impact. Neural codecs are potentially also better suited for new data formats, such as light fields for AR/VR applications, lidar data, or multi-view video. These technologies’ fast evolution will demand rapid prototyping, making learnable codecs appealing. Neural codecs also lack the common

block-coding visual artifacts and can be “supervised” to allocate more bits to certain features of interest. Neural codecs are more flexible than traditional codecs and can be optimized for superior perceptual quality or other custom metrics at much lower bitrates.

While the focus of the neural compression community has been largely on improving the tradeoff between bitrate and distortion (or perceptual quality), neural compression methods are currently 1-2 orders of magnitude slower than their classical counterparts. This makes their real-time deployment highly impractical, e.g., downloading and unpacking a large machine learning data set such as ImageNet or decoding video in real-time. By drawing on resource-efficient architectures, iterative inference, and new entropy coding schemes based on parsimonious models, we argue that the community should seek to maintain the strong performance of neural compression methods while increasing their runtime efficiency ideally by 1-2 orders of magnitude, removing one of the major obstacles from widespread deployment of neural codecs in the real world.

Many compression applications, such as video streaming on Youtube, impose strict runtime limitations upon decoding while allowing a much larger computational budget upon encoding. In “Improving Inference for Neural Image Compression” [Yang, Bamler, Mandt, NeurIPS 2020], we exploited this asymmetry by improving the encoding process using a larger computational budget while leaving the decoder untouched. To this end, we searched for an improved discrete latent representation at test time using an annealing scheme. This way, we obtained 15-20% rate savings without modifying the decoder. Our result suggests that iterative inference may achieve state-of-the-art compression performance with more lightweight decoders.

In contrast to most existing work focusing on architectural improvements in non-linear transform coding, we stress the importance of algorithmic advances that have broad applicability to existing methods, e.g., by exploiting the asymmetrical resource budgets for encoding and decoding via iterative inference and/or by developing new paradigms for entropy coding. Ways to improve the resource efficiency of neural codecs include drawing on lightweight decoders, iterative encoding at training time, and advances in parsimonious generative models. Our goals are to accelerate transform coding while also proposing new architectures for efficient entropy coding based on recent work on lossless compression with probabilistic circuits.

3.12 Informed Representation Learning with Deep Generative Models

Laura Manduchi (ETH Zürich, CH)

License © Creative Commons BY 4.0 International license

© Laura Manduchi

Joint work of Laura Manduchi, Thomas M. Sutter, Alain Ryser, Julia E. Vogt, Ricards Marcinkevics, Fabian Laumer, Joachim M. Buhmann, Kieran Chin-Cheong

One of the most popular frameworks to extract meaningful information from a vast amount of unlabeled datasets is representation learning. The latter should encode valuable information for downstream tasks, such as classification, regression, and visualization. However, there are many circumstances where purely data-driven approaches lead to unsatisfactory results that are inconsistent with the domain knowledge. On the other hand, deep generative models can encode physical laws and constraints into their generative process to obtain preferred representations of data, enabling exploratory analysis of complex data types. In this talk, I explored several approaches to incorporate domain knowledge, in the form of constraints,

probabilistic relations, and prior distributions, in VAEs for static and temporal data with a focus on clustering. First, I introduced two different gaussian mixture prior distributions used in VAEs to enforce a clustering structure in the latent space. The first one [1] employs a categorical prior distribution on the clusters, while the second one [2] uses a differentiable hypergeometric distribution to overcome the i.i.d. assumption of the input data. I then introduced the inclusion of domain knowledge in the form of probabilistic relations and survival information to obtain representations with a clear clustering structure for time series and survival data, using a mixture of Weibull distributions [3]. I then focused on instance-level constraints to guide the learning process toward a preferred configuration using high-dimensional data, using a Conditional Gaussian Mixture Model [4]. Last but not least, I showed how the proposed techniques can be applied to real-world medical applications with a focus on cardiology. In cardiovascular medicine, to correctly quantify cardiac function and diagnose dysfunction, expensive and time-consuming medical imaging methods are often required. This might lead to a lack of available diagnostic modalities and inadequate patient care. The use of machine learning models based on affordable and minimally invasive diagnostics, such as echocardiography, might serve as a valuable assistant tool to enhance health care for people with cardiovascular diseases. However, a lack of large labeled datasets in the medical field prevents the use of supervised deep learning techniques. Therefore, there is a need for informed representation learning algorithms to leverage prior information on unlabeled datasets of cardiac ultrasound videos. The learned representation can then be used to solve further downstream tasks, such as diagnosis, anomaly detection, and denoising [5], [6].

References

- 1 Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H. (2016). Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. International Joint Conference on Artificial Intelligence.
- 2 T. M Sutter, L. Manduchi, A. Ryser J. E. Vogt (2023). Learning Group Importance using the Differentiable Hypergeometric Distribution, International Conference on Learning Representations.
- 3 Manduchi, L., Marcinkevics, R., Massi, M.C., Gotta, V., Müller, T., Vasella, F., Neidert, M.C., Pfister, M., Vogt, J.E. (2022). A Deep Variational Approach to Clustering Survival Data. International Conference on Learning Representations.
- 4 Manduchi, L., Chin-Cheong, K., Michel, H., Wellmann, S., Vogt, J.E. (2021). Deep Conditional Gaussian Mixture Model for Constrained Clustering. Neural Information Processing Systems.
- 5 Ryser, A., Manduchi, L., Laumer, F., Michel, H., Wellmann, S., Vogt, J.E.. (2022). Anomaly Detection in Echocardiograms with Dynamic Variational Trajectory Models. Proceedings of the 7th Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research.
- 6 Laumer, F., Amrani, M., Manduchi, L., Beuret, A., Rubi, L., Dubatovka, A., Matter, C.M., Buhmann, J.M. (2022). Weakly supervised inference of personalized heart meshes based on echocardiography videos. Medical image analysis.

3.13 Towards Anytime Computation in Deep Architectures

Eric Nalisnick (University of Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Eric Nalisnick

Joint work of Eric Nalisnick, Allingham, James

Main reference James U. Allingham, Eric Nalisnick: “A Product of Experts Approach to Early-Exit Ensembles”.
Workshop on Dynamic Neural Networks at ICML, 2022.

URL https://dynn-icml2022.github.io/papers/paper_10.pdf

Predictive models often need to be evaluated in dynamic, uncertain computational conditions. For example, a model deployed to a mobile device should be useful, despite a lack of computational power. Moreover, the same model ideally should provide better performance on devices with richer computational resources. Hence, there is a need for “early-exit” architectures that allow computation to be halted early, before the model has run to completion. Current architectures provide little-to-no strong guarantees about how these intermediate predictions relate to the final prediction produced by the full model.

In this talk, I describe a construction that provide one such guarantee. Specifically, we guarantee that the probability of the mode under the full model monotonically increases in the intermediate solutions as more computation is done. This is achieved by a product-of-experts approach, as its predictive distribution takes the form of an intersection of the experts. We demonstrate that this architecture can be realized for both real-valued regression and multi-class classification.

3.14 The Future (R)evolution of Generative AI

Björn Ommer (LMU München, DE)

License © Creative Commons BY 4.0 International license
© Björn Ommer

Main reference Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer: “High-Resolution Image Synthesis with Latent Diffusion Models”, in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 10674–10685, IEEE, 2022.

URL <https://doi.org/10.1109/CVPR52688.2022.01042>

Recently, deep generative modeling has become the most prominent paradigm for learning powerful representations of our (visual) world and for generating novel samples thereof. Consequently, this has already become the main building block for numerous algorithms and practical applications. This talk will contrast the most commonly used generative models to date with a particular focus on denoising diffusion probabilistic models, the core of the currently leading approaches to visual synthesis. Despite their enormous potential, these models come with their own specific limitations. We will then discuss a solution, latent diffusion models, a.k.a. “Stable Diffusion”, that significantly improves the efficiency of diffusion models. Now billions of training samples can be summarized in compact representations of just a few gigabytes so that the approach runs on consumer GPUs. We will then discuss recent extensions that cast an interesting perspective on future generative models. In particular, retrieval augmentation during inference promises to significantly reduce model sizes by having powerful likelihood models focus on the composition of a scene rather than learning the training data. We will then highlight key aspects in which generative modeling will change in the future.

3.15 Where to Diffuse, how to diffuse, and how to get back?

Rajesh Ranganath (NYU Courant Institute of Mathematical Science, US)

License © Creative Commons BY 4.0 International license
© Rajesh Ranganath

Main reference Raghav Singhal, Mark Goldstein, Rajesh Ranganath: “Where to Diffuse, How to Diffuse, and How to Get Back: Automated Learning for Multivariate Diffusions”, CoRR, Vol. abs/2302.07261, 2023.

URL <https://doi.org/10.48550/arXiv.2302.07261>

Generative models have been making large leaps in both quantitative fidelity and qualitative appeal. One class of models that has been a driving force behind these leaps is diffusion-based generative models. Diffusion-based generative models, or diffusions models for short, work by first corrupting data towards a known, fixed stationary distribution and training a model to undo those corruptions, and thus providing a means to generate data by uncorrupting a sample from the stationary distribution. The choice of corruption or inference process affects both likelihoods and sample quality. For example, it has been shown that extending the inference diffusion with auxiliary variables, making them multivariate, leads to improved sample quality. However, deriving training algorithms for each new inference diffusion is onerous requiring manually deriving stationary distributions and transition kernels. To simplify the training, we provide a recipe for likelihood training of multivariate diffusion models. In the first step, we derive a lower bound on the likelihood. Next, we show how the terms in the lower bound can be automatically computed and show how to parametrize inference diffusions using results from Markov chain Monte Carlo to target a specific stationary distribution. We study several different inference diffusions and demonstrate how to learn and the value of learning inference diffusions.

3.16 Universal Critics

Lucas Theis (Google – London, GB)

License © Creative Commons BY 4.0 International license
© Lucas Theis

What distinguishes a realistic image from an unrealistic image? Despite tremendous progress in our ability to generate realistic images, we still lack functions that can reliably detect artifacts in images. Such a function would be of great interest in a variety of applications such as outlier detection, perceptual quality evaluation, neural compression, or neural rendering. The field of algorithmic probability provides many insights on a closely related question; namely, when is data a plausible sample from a distribution P ? However, these results are not widely known in the machine learning community. In this short presentation, I will discuss how Kolmogorov complexity can inspire “universal critics” – functions that are able to detect unrealistic images without being trained on corrupted data.

3.17 Getting the most ****out**** of your representations

Karen Ullrich (Meta – New York, US)

License © Creative Commons BY 4.0 International license
© Karen Ullrich

Joint work of Karen Ullrich, Yangjun Ruan, Daniel Severo, James Townsend, Ashish Khisti, Arnaud Doucet, Alireza Makhzani, Yann Dubois, Benjamin Bloem-Reddy, Chris J Maddison

Main reference Yangjun Ruan, Karen Ullrich, Daniel Severo, James Townsend, Ashish Khisti, Arnaud Doucet, Alireza Makhzani, Chris J. Maddison: “Improving Lossless Compression Rates via Monte Carlo Bits-Back Coding”, in Proc. of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, Vol. 139, pp. 9136–9147, PMLR, 2021.

URL <http://proceedings.mlr.press/v139/ruan21a.html>

Main reference Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, Chris J Maddison: “Lossy Compression for Lossless Prediction”, in Proc. of the Advances in Neural Information Processing Systems, Vol. 34, pp. 14014–14028, Curran Associates, Inc., 2021.

URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7535bbb91c8fde347ad861f293126633-Paper.pdf

The goal of source compression is to map any outcome of a discrete random variable $x \sim p_d(x)$ in a finite symbol space $x \in S$ to its shortest possible binary representation. Given a tractable model probability mass function (PMF) $p(x)$ that approximates $p_d(x)$, entropy coders provide such an optimal mapping. As a result, the task of source compression is simplified to identifying a good model PMF for the data at hand.

Even though the setup as described is the most commonly used one, there are restrictions to it. Entropy coders can only process one dimensional variables and process them sequentially. Hence the structure of the entropy coder implies a sequential structure of the data. This is a problem when compressing sets instead of sequences. In the first part of the talk, I present an optimal codec for sets [1]. The problem we encounter for sets can be generalized for many other structural priors in data. In the second part of the talk I thus investigate the problem. We generalize rate distortion theory for structural data priors and develop a strategy to learn codecs for this data [2].

References

- 1 D. Severo, J. Townsend, A. Khisti, A. Makhzani and K. Ullrich. “Compressing Multisets with Large Alphabets.” IEEE Journal on Selected Areas in Information Theory 2023.
- 2 Dubois, Y., Bloem-Reddy, B., Maddison, C.J. “Lossy Compression for Lossless Prediction.” Neural Information Processing Systems 2021.

3.18 Interventional causal representation learning with deep generative models

Yixin Wang (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 4.0 International license
© Yixin Wang

Joint work of Yixin Wang, Kartik Ahuja, Divyat Mahajan, Yoshua Bengio, Michael I Jordan

Main reference Yixin Wang, Michael I. Jordan: “Desiderata for Representation Learning: A Causal Perspective”, CoRR, Vol. abs/2109.03795, 2021.

URL <https://arxiv.org/abs/2109.03795>

Main reference Kartik Ahuja, Yixin Wang, Divyat Mahajan, Yoshua Bengio: “Interventional Causal Representation Learning”, CoRR, Vol. abs/2209.11924, 2022.

URL <https://doi.org/10.48550/arXiv.2209.11924>

Causal representation learning seeks to extract high-level latent factors from low-level sensory data. Most existing methods rely on fitting deep generative model to observational data, leveraging structural assumptions (e.g., conditional independence) to identify the latent

factors. However, interventional data is prevalent across applications. Can interventional data facilitate causal representation learning? We explore this question in this talk. The key observation is that interventional data often carries geometric signatures of the latent factors’ support (i.e. what values each latent can possibly take). For example, when the latent factors are causally connected, interventions can break the dependency between the intervened latents’ support and their ancestors’. Leveraging this fact, we prove that the latent causal factors can be identified up to permutation and scaling given data from perfect do interventions. Moreover, we can achieve block affine identification, namely the estimated latent factors are only entangled with a few other latents if we have access to data from imperfect interventions. These results highlight the unique power of interventional data in causal representation learning; they can enable provable identification of latent factors without any assumptions about their distributions or dependency structure.

3.19 Assaying Out-Of-Distribution Generalization in Transfer Learning

Florian Wenzel (Amazon Web Services – Tübingen, DE)

License  Creative Commons BY 4.0 International license
© Florian Wenzel

Joint work of Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, Francesco Locatello


Main reference Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, Francesco Locatello: “Assaying Out-Of-Distribution Generalization in Transfer Learning”, CoRR, Vol. abs/2207.09239, 2022.

URL <https://doi.org/10.48550/arXiv.2207.09239>

Since out-of-distribution generalization is a generally ill-posed problem, various proxy targets (e.g., calibration, adversarial robustness, algorithmic corruptions, invariance across shifts) were studied across different research programs resulting in different recommendations. While sharing the same aspirational goal, these approaches have never been tested under the same experimental conditions on real data. In this paper, we take a unified view of previous work, highlighting message discrepancies that we address empirically, and providing recommendations on how to measure the robustness of a model and how to improve it. To this end, we collect 172 publicly available dataset pairs for training and out-of-distribution evaluation of accuracy, calibration error, adversarial attacks, environment invariance, and synthetic corruptions. We fine-tune over 31k networks, from nine different architectures in the many- and few-shot setting. Our findings confirm that in- and out-of-distribution accuracies tend to increase jointly, but show that their relation is largely dataset-dependent, and in general more nuanced and more complex than posited by previous, smaller scale studies.

3.20 Trading Information between Latents in Hierarchical Variational Autoencoders

Robert Bamler (Universität Tübingen, DE)

License  Creative Commons BY 4.0 International license

© Robert Bamler

Joint work of Robert Bamler, Tim Xiao

Motivated by the recent success stories of generative modeling with diffusion models, we reconsider the training objective of hierarchical variational autoencoders (VAEs). By keeping the so-called forward process (from data to latent representations) fixed and recurrent, diffusion models are able to efficiently scale up training to very deep probabilistic models with many layers of latents, leading to impressive generative modeling performances. However, it is not so clear how to use diffusion models for applications that make use of the forward process, such as representation learning or data reconstruction tasks, as the fixed forward diffusion process progressively reduces mutual information with the original data up to the point of almost no correlation. For these types of applications, VAEs with their learned inference models are often a more natural choice of model architecture. Motivated by the empirical observation that deep hierarchies of layers of latent variables are crucial to the performance of diffusion models, we reconsider the trade-off between reconstruction error (“distortion”) and information content of the latents (“rate”) in hierarchical VAEs, i.e., VAEs with more than one layer of latents.

We observe first that the general rate/distortion trade-off of beta-VAEs [Alemi et al., ICML 2018] can be refined by splitting up the rate term into a sum of contributions from each layer of latents. Importantly, however, this separation is only possible if the inference model proceeds in the same direction as the generative model, i.e., opposite to the direction of the otherwise analogous forward process in diffusion models. Splitting the rate into layer-wise contributions allows practitioners to control the information content of each layer individually by introducing individual layer-wise Lagrange multipliers (“beta hyperparameters”). We argue that this increased control is useful in practice by grouping application domains of (hierarchical) VAEs into three categories depending on whether they use (i) only the generative model (generative tasks), (ii) only the inference model (representation learning tasks), or (iii) both (reconstruction tasks). We show both by deriving theoretical performance bounds, as well as by large-scale empirical evaluations that the optimal distribution of rate between layers of latents is different for the three categories of applications, and we provide practical guidance for choosing reasonable values for the beta hyperparameters for each category.

The main open question that we will consider in the future is if the proposed hierarchical rate/distortion theory can be used to train VAEs where the size of the latent representation does not necessarily match between the inference and the generative process. Here, the highest-level latents could represent the length of the next lower-level latent representation. Such a variable-length information bottleneck would allow training VAEs with transformer architectures for text, e.g., building on the work by Henderson and Fehr (arXiv:2207.13529). Treating the length as a (higher-level) latent variable would allow training scenarios where the length of the reconstructed text does not necessarily match the length of the original text. The hope is that this would allow a VAE to more freely rephrase text, and that the individual beta-hyperparameters would allow controlling the length and the diversity of reconstructed text separately.

3.21 Challenges in Generative Language Modeling

Alexander Rush (Cornell University – Ithaca, US)

License  Creative Commons BY 4.0 International license
© Alexander Rush

My talk will argue that Generative models are the defining element of modern NLP. I will describe some of the recent usage of generative models for NLP. Specifically the now popular knowledge that they are extremely impressive tools that are central to the field. Understanding that this is different in spirit than the main focus of the seminar, I will try to start a discussion as to why NLP generative models are somehow less powerful than other generative approaches. Given this context, I will describe some of the remaining modeling challenges in using these systems. Specifically, language models have demonstrated the ability to generate highly fluent text; however, they still require additional scaffolding to maintain coherent high-level structure (e.g., story progression). Using the model criticism in latent space we can evaluate the high-level structure by comparing distributions between real and generated data in a latent space obtained according to an assumptive generative process. Different generative processes identify specific failure modes of the underlying model. We perform experiments on three representative aspects of high-level discourse – coherence, coreference, and topicality – and find that transformer-based language models are able to capture topical structures but have a harder time maintaining structural coherence or modeling coreference structures. Based on these conclusions, I pose questions about how we might update our models for language and ask whether richer generative processes might better capture some aspects of language current systems are missing.

3.22 Fun with Foundation Models and Amortized Inference


Frank Wood (University of British Columbia – Vancouver, CA)

License  Creative Commons BY 4.0 International license
© Frank Wood

In this talk, I will discuss the work of my UBC PLAI group and spin-out Inverted AI on foundation models of behavior, images, and video. In particular, I will talk about ways to get such models to “do what you want them to do” via amortized inference after they have been trained. I will spend most of my time introducing and talking about ITRA, a model I think can become the GPT of behavior, and how we use inference, including a novel algorithm called “critic SMC,” alongside reinforcement learning as inference techniques, to “tune” ITRA to stay on the road in new places, not collide with other agents, and produce trajectories that are achievable in the dynamics sense by specific vehicle classes. I will then discuss very related techniques for amortized conditional inference in image and video generative models, work that has led to state-of-the-art inpainting and conditional image generation results requiring no fine-tuning of a pre-trained VAE image model and stunning recent results on very long duration photorealistic video generation arising from meta-learning a flexible conditioning DDPM-based video generative model architecture.

3.23 Languages for the Next 700 Application Domains in AI

Jan-Willem van de Meent (University of Amsterdam, NL)


License  Creative Commons BY 4.0 International license
© Jan-Willem van de Meent

This will be a talk about where AI has arrived today, where we could be going in the next few years, and the role that probabilistic approaches to AI have to play in these developments. I will discuss where I see opportunities in applications of AI to computational design in the physical sciences, and I will discuss how programming language design can help realize these opportunities, with particular attention to our recent work on languages for inference programming, which opens up opportunities for new model and inference designs, both in the context of simulation-based inference and in the context of deep generative models.

4 Working groups

4.1 Continual learning of deep generative models

Sophie Fellenz (RPTU – Kaiserslautern, DE), Sina Däubener (Ruhr-Universität Bochum, DE), Gerard de Melo (Hasso-Plattner-Institut, Universität Potsdam, DE), Florian Wenzel (Amazon Web Services – Tübingen, DE), and Frank Wood (University of British Columbia – Vancouver, CA)

License  Creative Commons BY 4.0 International license
© Sophie Fellenz, Sina Däubener, Gerard de Melo, Florian Wenzel, and Frank Wood

Given that foundation models are increasing in size, the training time for these models is also increasing, which makes frequent retraining impractical. A more desirable alternative would be to continually update existing models with new data. This way we could build agents that learn continuously (life-long learning) and make learning more efficient. The main challenge we identified here is “catastrophic forgetting”. How can we make sure that the model incorporates new information without forgetting what it already knows? Many techniques have been proposed in the supervised setting where new labels can be added over time, but we posit that in order to solve continual learning in the supervised setting, it first needs to be solved in the unsupervised setting. We discussed techniques such as fine-tuning, functional regularization, context augmentation, storing part of the data (core sets) or storing a part of learned parameters. None of these are satisfactory solutions as they cannot guarantee to prevent catastrophic forgetting and are hard to optimize or control in practice. Sequential Monte Carlo or streaming variational Bayes are theoretically possible but do not scale in practice. We also discussed hybrid approaches that learn representations using neural networks and apply a Kalman Filter or similar on the condensed representations. As interesting directions we furthermore identified active learning on prompts and adversarial interactions.

4.2 Priors in deep generative modeling

Vincent Fortuin (University of Cambridge, GB), Thomas Gärtner (Technische Universität Wien, AT), Matthias Kirchler (Hasso-Plattner-Institut, Universität Potsdam, DE), and Eric Nalisnick (University of Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Vincent Fortuin, Thomas Gärtner, Matthias Kirchler, and Eric Nalisnick

This group discussed the problem of how to elicit priors for deep generative models from domain experts and how to encode them in the model. The running example was the problem of learning a generative model for antibiotic drugs from a small set of existing molecules (on the order of a few hundred). Standard deep generative modeling would probably not be data-efficient enough to learn a sufficiently expressive model from such a small dataset, but the hope would be to use prior knowledge from domain experts, such as chemists. The problem is that this knowledge is often rather vague in the chemist’s mind, for instance, some rough intuition for what kind of functional groups typical antibiotics should have, or how large or aromatic they are. One promising idea to elicit this prior knowledge from the expert is to show them molecules and then query their beliefs about them, either by asking them whether this looks like an antibiotic (binary feedback) or how much it looks like an antibiotic (continuous feedback). The molecules to show them can be samples from the model before training, so essentially from the prior predictive, which would then allow directly tuning the prior of the model based on the feedback. Alternatively, we can use a large dataset of unlabeled chemicals to collect feedback on and then use that to distill a prior. Both of these approaches have the disadvantage that most of the molecules we would show the expert would probably not look like antibiotics, so we would waste a lot of their time for not a lot of information. One would probably need to use some acquisition function, similar to active learning, to try and only ask about the most informative compounds in each iteration. We also discussed in this vein that the process could be made easier by treating the expert knowledge as a likelihood instead of a prior, that is, first training the generative model based on the small dataset and then finetuning its generations with the human feedback, which would hopefully create more interesting structures than the prior predictive. Moreover, if we ask the expert to make changes to the generated molecules to make them more drug-like, we could use a distribution over these changes as a prior for the score function in score-based models such as diffusion models. We also discussed that we could try to use weak supervision signals such as human-defined similarity measures or hand-designed features to build a classifier from the human feedback and then train the generative model with classifier-guidance instead of a proper prior. As a mechanism to incorporate the prior knowledge into the model, except for the aforementioned proper prior distributions or classifier-guidance, we also discussed rejection sampling, importance sampling, and probabilistic circuits. Overall, we concluded that directly incorporating vague human knowledge into a proper prior distribution seems hard and that approaches based on iterative human feedback are probably more promising.

4.3 The role of domain knowledge in deep generative models

Vincent Fortuin (University of Cambridge, GB), Thomas Gärtner (Technische Universität Wien, AT), Matthias Kirchler (Hasso-Plattner-Institut, Universität Potsdam, DE), Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE), Laura Manduchi (ETH Zürich, CH), Guy Van den Broeck (UCLA, US), Julia Vogt (ETH Zürich, CH), and Florian Wenzel (Amazon Web Services – Tübingen, DE)

License © Creative Commons BY 4.0 International license
 © Vincent Fortuin, Thomas Gärtner, Matthias Kirchler, Christoph Lippert, Laura Manduchi, Guy Van den Broeck, Julia Vogt, and Florian Wenzel

In this working group we discussed different types of domain knowledge and how it can be incorporated into deep generative models. Firstly, we discussed why it can be useful to incorporate domain knowledge and agreed that it might improve data-efficiency, enable extrapolation beyond the training dataset, and increase trustworthiness of the model. Secondly, we discussed different kinds of domain knowledge and how they could respectively be implemented in deep generative models. Knowledge about causality can be incorporated using causal graphs or structural equation models in some latent space, which then implies disentanglement of the latent factors corresponding to causal factors. More generally, known probabilistic relationships can be represented in the form of prior distributions, either in the latent space or the data space directly. Physical constraints, invariances, and symmetries can often be incorporated directly through the choice of model architecture, for instance, CNNs in the case of images. Facts about the world can be incorporated via knowledge graphs or database retrieval mechanisms, while logical statements can be incorporated through fuzzy or probabilistic logic. Finally, ontologies can be incorporated via hierarchical modeling. Overall, we concluded that one of the main challenges is still to design a latent space in which representations carry semantic meaning, since many of these types of domain knowledge would need to be incorporated into such a latent space. This is loosely related to the problem of symbol grounding from continuous distributed representations.

4.4 Anomaly detection using Kolmogorov complexities

Marius Kloft (RPTU – Kaiserslautern, DE), Asja Fischer (Ruhr-Universität Bochum, DE), and Lucas Theis (Google – London, GB)

License © Creative Commons BY 4.0 International license
 © Marius Kloft, Asja Fischer, and Lucas Theis

Ideally, one would threshold the log-(pseudo)likelihood ratio $s = \log(p/q)$ of the distribution of the normal data p and the distribution of the anomalous data q for provably optimal anomaly detection. In practice, q is unknown, and one resorts to thresholding $\log(p)$. Steinwart (2005) showed that this could be equivalent to thresholding $\log(p/q)$ where q is a uniform distribution of anomalies. In practice, however, it has been observed that anomalies typically are “simpler” (easier to compress) – a phenomenon known as “Occam’s razor”. We propose to replace $\log(q)$ in s by an approximation of k , the Kolmogorov complexity, which – intuitively speaking – measures the likeliness of an instance occurring in nature. Furthermore, as an extension, one could integrate the Kolmogorov complexity with an estimate of q based on some observed anomalies.

Participants

- Robert Bamler
Universität Tübingen, DE
- Ryan Cotterell
ETH Zürich, CH
- Sina Däubener
Ruhr-Universität Bochum, DE
- Gerard de Melo
Hasso-Plattner-Institut,
Universität Potsdam, DE
- Sophie Fellenz
RPTU – Kaiserslautern, DE
- Asja Fischer
Ruhr-Universität Bochum, DE
- Vincent Fortuin
University of Cambridge, GB
- Thomas Gärtner
Technische Universität Wien, AT
- Matthias Kirchler
Hasso-Plattner-Institut,
Universität Potsdam, DE
- Marius Kloft
RPTU – Kaiserslautern, DE
- Yingzhen Li
Imperial College London, GB
- Christoph Lippert
Hasso-Plattner-Institut,
Universität Potsdam, DE
- Stephan Mandt
University of California –
Irvine, US
- Laura Manduchi
ETH Zürich, CH
- Eric Nalisnick
University of Amsterdam, NL
- Björn Ommer
LMU München, DE
- Rajesh Ranganath
NYU Courant Institute of
Mathematical Science, US
- Maja Rudolph
Bosch Center for AI –
Pittsburgh, US
- Alexander Rush
Cornell University – Ithaca, US
- Lucas Theis
Google – London, GB
- Karen Ullrich
Meta – New York, US
- Jan-Willem van de Meent
University of Amsterdam, NL
- Guy Van den Broeck
UCLA, US
- Julia Vogt
ETH Zürich, CH
- Yixin Wang
University of Michigan –
Ann Arbor, US
- Florian Wenzel
Amazon Web Services –
Tübingen, DE
- Frank Wood
University of British Columbia –
Vancouver, CA

