ETH zürich

Distributed Computing Group

Seminar in Deep Neural Networks 2024

Benjamin Estermann, Florian Grötschla

20th of February 2024

Introduce yourself!

- Name
- Degree, Background in Machine Learning (theoretical and/or practical)
- What are your expectations for the seminar?
- What do you want to learn?

Supervisors


Peter


Benjamin


Frédéric


Florian


Andreas

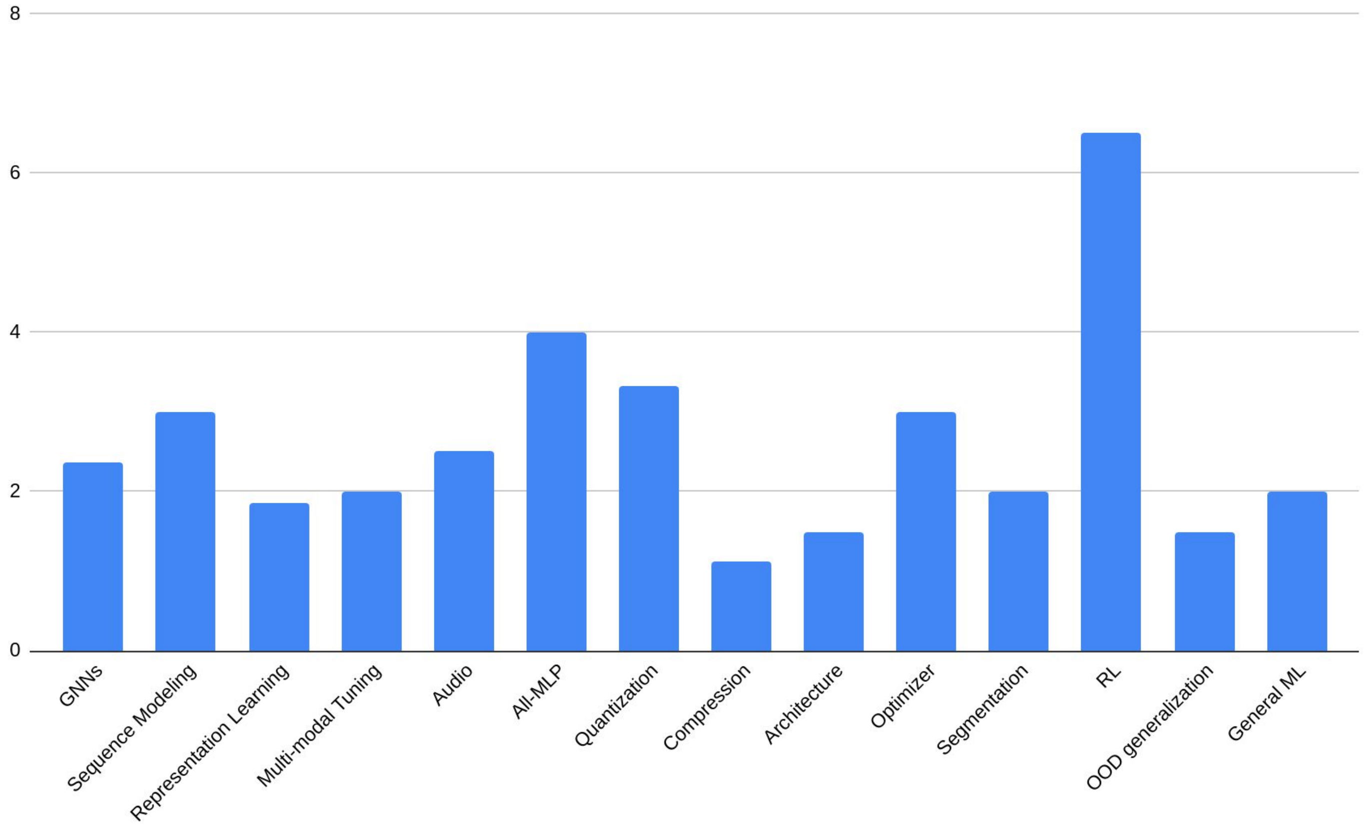

Till


Luca


Mattia


Joël

Topic overview



Schedule

Date	Presenter	Title	Mentor	Slides
February 20	Benjamin Estermann	Introduction to Scientific Presentations	-	TBA
February 27	Dennis Jüni Denis Tarasov	Simple and Controllable Music Generation Direct Preference Optimization: Your Language Model is Secretly a Reward Model	Luca Lanzendörfer Frédéric Berdoz	TBA TBA
March 5	Jiaqing Xie Ning Wang	Graph Inductive Biases in Transformers without Message Passing Disentanglement with Biological Constraints: A Theory of Functional Cell Types	Florian Grötschla Benjamin Estermann	TBA TBA
March 12	Kim Yumi Davide Guidobene	AudioLDM: Text-to-Audio Generation with Latent Diffusion Models Maximally Expressive GNNs for Outerplanar Graphs	Luca Lanzendörfer Florian Grötschla	TBA TBA
March 19	Guiv Farmanfarmaian Eric Nothum	Agree to Disagree: Diversity through Disagreement for Better Transferability Mamba: Linear-Time Sequence Modeling with Selective State Spaces	Frédéric Berdoz Mattia Segu	TBA TBA
March 26	Pyrros Koussios Zixuan Chen	Siamese Masked Autoencoders Controlling Rate, Distortion, and Realism: Towards a Single Comprehensive Neural Image Compression Model	Mattia Segu Till Aczel	TBA TBA
April 02	-	Easter Break	-	-

...

How to structure your talk

Introduction	Previous work	Contribution of the Paper

Presentation Style

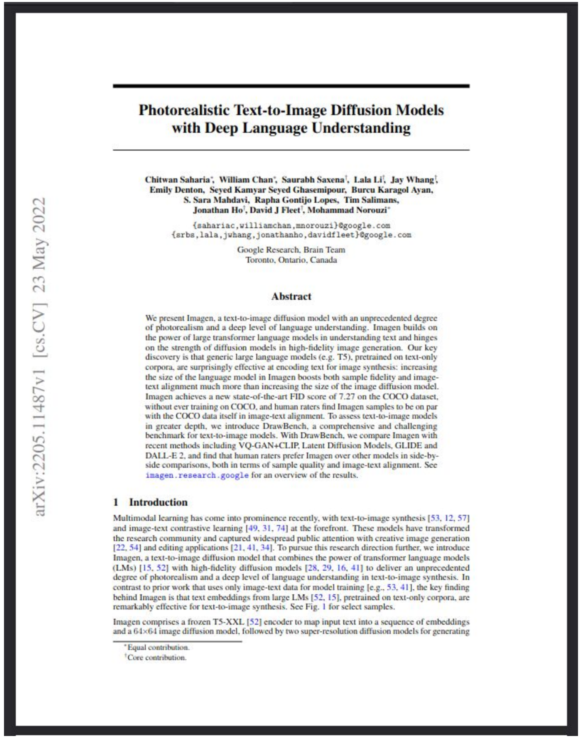
Great Scientific Presentations
by Roger Wattenhofer

Let's start with some general points

- If you can pull off to give a talk **without slides**, you will be admired! Don't hesitate to use the blackboard (if one exists) for some parts of your talk. That said, slides do help the rest of us. On the second page of this document is some advice specifically for slides.
- Do *not* explain **every detail** of the work. Give an exciting talk, not a talk that lists everything that was done.
- Your talk must have parts that can be fully understood by the audience, parts where the audience learns something. Maybe (hopefully) there is not enough time to show every detail? Or maybe some details are just tedious, but not really interesting? It is okay to **sketch** some parts only. If some aspect is only presented on a high level, make sure that the audience understands that you simplified for the sake of the presentation.
- Some students have started giving management style talks when presenting their work! This is of course a big no-no when it comes to science and technology. You definitely must present the most **interesting technical and theoretical aspects** of the work!
- What are the **motivating examples**? What are the examples that render a naive approach impossible? Why does the model need this strange additional assumption? Where is the struggle and why? What is the most surprising part of the work? Your talk should be full of these examples. Instead of explaining a dry model, explain a problem in a natural way, and then explain the model along with examples.
- The ultimate example is the **demo**. Most audiences love a great demo. Don't wait with your demo until the end of your talk. A demo could also be at the very beginning of your talk, or in the middle, or throughout your talk.
- **Know your audience**: A lecture to undergrad students is different from a conference talk. Is your audience waiting for your talk (job interview presentation), or is it sitting there for three days already, listening to one mediocre talk after the other, desperate for something different?
- Try to **keep your audience** throughout your talk. It may be okay to lose a certain fraction of the audience from time to time (for a bit), it is *not* okay to lose 50% of the audience during 50% of the talk.
- Use **metaphors**. A metaphor is a glorious thing.
- If possible, **interact** with your audience.
- Have a good standing **posture**.
- Be **on time**. Actually, don't mind finishing 1' early. Nobody is going to be mad.
- Be funny, be deep. **Don't be boring**!

https://disco.ethz.ch/courses/fs24/seminar/GreatScientificPresentations.pdf

How to approach your paper



arXiv:2205.11487v1 [cs.CV] 23 May 2022

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac,williamchan,mnorouzi}@google.com
{srbs,lala,jwhang,jonathanho,davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

Abstract

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, GLIDE and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment. See imagen.research.google for an overview of the results.

1 Introduction

Multimodal learning has come into prominence recently, with text-to-image synthesis [53, 12, 57] and image-text contrastive learning [49, 31, 74] at the forefront. These models have transformed the research community and captured widespread public attention with creative image generation [22, 54] and editing applications [21, 41, 34]. To pursue this research direction further, we introduce Imagen, a text-to-image diffusion model that combines the power of transformer language models (LMs) [15, 52] with high-fidelity diffusion models [28, 29, 16, 41] to deliver an unprecedented degree of photorealism and a deep level of language understanding in text-to-image synthesis. In contrast to prior work that uses only image-text data for model training [e.g., 53, 41], the key finding behind Imagen is that text embeddings from large LMs [52, 15], pretrained on text-only corpora, are remarkably effective for text-to-image synthesis. See Fig. 1 for select samples.

Imagen comprises a frozen T5-XXL [52] encoder to map input text into a sequence of embeddings and a 64×64 image diffusion model, followed by two super-resolution diffusion models for generating

*Equal contribution.
†Core contribution.

OpenReview.net

ICLR

NEURAL INFORMATION PROCESSING SYSTEMS

Admin stuff