Nikhil Kapila

+971503564790 | nkapila.me | github | nkapila6@gatech.edu

Research Interests

My research interests lie in optimization of deep learning architectures,multimodal image and text models, and generative image/speech models.

Optimization

In optimization, I would like to look into techniques like knowledge distillation, neural architectural search and pruning to create smaller models that outperform their larger counterparts. I tried to this recently with my recent project `CNNtention` [1], we augmented attention layers in a deep ResNet model to see if refined feature maps through attention yield any benefit. For upcoming work, in my personal downtime, I plan to look into knowledge distillation techniques, i.e. some form of intermediate feature map alignment using a bigger model.

Apart from architectural changes, compute efficiency can be improved through alternative methods. It could be something simple like decomposing the attention matrix (bilinear attention) by approximating the full `Wx` to save on compute. Another example is replacing traditional matmuls with Hadamard products as seen in certain GRU implementations [2].

Multimodal and Generative

As I dive more into the world of deep learning, I am able to recognize many similarities across modalities. I realize a lot of the ideas from image could be transferred to other modalities such as speech/text with ease.

One example is the case of using `Neural Style Transfer for Images` [3] and the case of using specific loss functions to dictate these changes [4]. I see a lot of similarities in the audio/speech modality where one could perform voice cloning by performing a similar style transfer approach in audio [5]. Of course, there are many newer pre-trained few/zero-shot speech models that exist now.

It's fascinating to observe how architectures with similar design principles can operate across different modalities by learning appropriate latent representations, as demonstrated by contrastive learning approaches like CLIP that create aligned embedding spaces. The recent emergence of foundation models has further highlighted how architectural decisions can enable powerful cross-modal capabilities, allowing systems to reason across text, images, audio, and other modalities with increasing coherence and sophistication.

Research Trajectory

I am due to complete my MS in Fall of 2025. Currently I am working on a few projects as listed below:
Coursework

- ICD9 disease prediction from unstructured clinical notes of MIMIC-III. Using the `FarSight` paper as our base [6].

Personal Projects: Projects I am trying to accomplish in my downtime.

- **Implementing GANs and VAEs**: I am trying to investigate latent space properties that enable effective generative modeling.
- **Implementation of models such as GPT2**: To get a better understanding of what goes into pre-training large language models.

Nikhil Kapila

+971503564790 | nkapila.me | github | nkapila6@gatech.edu

References

[1] Nikhil Kapila, Julian Glattki, and Tejas Rathi, "CNNtention: Can CNNs learn better with Attention?," Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.11657v3

[2] Rui-Jie Zhu *et al.*, "Scalable MatMul-free Language Modeling," Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.02528

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "A Neural Algorithm of Artistic Style," Aug. 2015. [Online]. Available: https://arxiv.org/abs/1508.06576v2

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," Sep. 2023. [Online]. Available: https://arxiv.org/abs/1603.08155

[5] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao, "GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech," Oct. 2022.

[6] Tushaar Gangavarapu, G. Krishnan, S. SowmyaKamath, and Jayakumar Jeganathan, "FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes," Jul. 2021.