Statement of Purpose

of James Baker (Ph.D. applicant for Fall 2024)

Recent breakthroughs have revolutionized how computers create, augment, and understand unstructured data like images, text, and video. Pursuing a Ph.D. in computer science at MIT is the best way to explore my passion for using AI for creative work, like writing stories and making art. I want to democratize the tools for computational artistic expression so hobbyists, amateurs, and professionals can all realize their visions. That could entail improving semantic imaging for generative models, annotating datasets with new types of human feedback, and building new specialized architectures for specific creative domains. Given my background in Natural Language Processing and Computer Vision, I have the experience necessary to succeed.

My initial exposure to Natural Language Processing began with an honors thesis project under Dr. Eugene White, predicting stock prices using text data scraped from social media. I devised a novel approach (Baker (2023b)) to use k-means (MacQueen et al. (1967)) to cluster the embeddings of each comment and used the daily frequency of comments classified as belonging to each cluster as a factor into a traditional Capital Asset Pricing Model (Fama and French (2004)). This demonstrates my ability to develop and apply new ideas in unconventional domains. Beginning in July 2023, I was a Natural Language Processing engineer at a startup, supplementing a conversational agent with fine-tuning and Retrieval Augmented Generation (Lewis et al. (2020)). This exposed me to some of the problems with language models, such as hallucinations and the reversal curse (Berglund et al. (2023)), that need to be solved for creative works like stories and poems.

My first experience with Computer Vision was at the University of Houston. Under Dr. Ioannis Pavlidis, I architected a model using visual and channel-wise attention for facial landmark detection in thermal images, a previously underexplored domain. At the end of the summer, I presented my work alongside other research assistants, winning the award for second best. The following semester at Rutgers, I did an independent study under Dr. Ahmed Elgammal, where I designed ARTEMIS (Baker (2023a)), a GAN (Goodfellow et al. (2014)) framework for abstract texture synthesis. My unique approach was that each discriminator was trained not on images but on the style and content features extracted from the images using a pre-trained VGG network (Simonyan and Zisserman (2015)). The project earned me the accolade of being named a Paul Robeson scholar.

There are a few problems I am interested in solving. First, semantic image editing is a promising line of work to align generative outputs with users' artistic vision. The advent of techniques to increase the context window of language models (Han et al. (2023); Xiao, Tian, Chen, Han, and Lewis (2023)) allows longer text prompts to guide image synthesis. New methods of prompt engineering and prefix tuning (Li and Liang (2021)) could further improve these models. Editing the latent space of diffusion models allows users to fine-tune generated artifacts (Haas, Huberman-Spiegelglas, Mulayoff, and Michaeli (2023)), but this technique relies on domain-specific pre-trained classifiers. I plan to use classifier-free guidance (Ho and Salimans (2022)) to allow this technique to generalize. Second, we need better empirical data on what humans define as creative. While there are many datasets of visual art (Liao, Li, Liu, and Keutzer (2022); Saleh and Elgammal (2015)) and literature (Fan, Lewis, and Dauphin (2018); Kobayashi (2018)), they could be supplemented by adding human ratings of novelty and utility, those being the considered the components of creativity (Cropley (2006)) or biological signals that correlate with subjective appreciation (Dmochowski et al. (2014); Dorr, Martinetz, Gegenfurtner, and Barth (2010)). Third, the unique structure of creative work can lend itself to specialized techniques to generate them. Given methods to turn stories into sequences of discrete events (Martin et al. (2018)), inverse reinforcement learning (Abbeel and Ng (2004)), or recommendation system-based techniques could be used to generate these sequences and, in turn, produce novel stories. Images used for artistic purposes (unlike medical images) have both style and content. Segmentation models (Kirillov et al. (2023)) can be used to find objects and locations, representing the content of an image separate from style. This opens the possibility of generation models that, given a source image, maximize exploration in content or style and minimize divergence from the source in the other when producing a target image.

I aim to continue research after my Ph.D., either in an academic or industry laboratory. To that end, MIT would be an excellent fit for me. While I would be open and honored to work with any of the faculty at MIT, there are a few that I am most compatible with. Fox Harrell and I are both fascinated with exploring how machines can cooperate with and amplify human imagination, serving as tools for enriching creative pursuits. Aude Oliva's research covers computational models of human cognitive processes, encompassing aspects of artistic creativity and subjective reactions to visual stimuli—areas particularly relevant to my interest in AI and art.

References

Abbeel, P., & Ng, A. (2004, 09). Apprenticeship learning via inverse reinforcement learning. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*. doi: 10.1007/978-0-387-30164-8_417

Baker, J. (2023a). *Artemis: Using gans with multiple discriminators to generate art.*

Baker, J. (2023b). *In the red(dit): Social media and stock prices.*

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). *The reversal curse: Llms trained on "a is b" fail to learn "b is a".*

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391-404. Retrieved from `https://doi.org/10.1207/s15326934crj1803_13` doi: 10.1207/s15326934crj1803_13

Dmochowski, J. P., Bezdek, M. A., Abelson, B., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*, *5*. Retrieved from `https://api.semanticscholar.org/CorpusID:9267175`

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, *10 10*, 28. Retrieved from `https://api.semanticscholar.org/CorpusID:6650667`

Fama, E. F., & French, K. R. (2004, September). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, *18*(3), 25-46. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/0895330042162430` doi: 10.1257/0895330042162430

Fan, A., Lewis, M., & Dauphin, Y. (2018). *Hierarchical neural story generation.*

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.*

Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., & Michaeli, T. (2023). *Discovering interpretable directions in the semantic latent space of diffusion models.*

Han, C., Wang, Q., Xiong, W., Chen, Y., Ji, H., & Wang, S. (2023). *Lm-infinite: Simple on-the-fly length generalization for large language models.*

Ho, J., & Salimans, T. (2022). *Classifier-free diffusion guidance.*

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R. (2023). Segment anything. *arXiv:2304.02643*.

Kobayashi, S. (2018). *Homemade bookcorpus.* `https://github.com/soskek/bookcorpus`.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, *abs/2005.11401*. Retrieved from `https://arxiv.org/abs/2005.11401`

Li, X. L., & Liang, P. (2021). *Prefix-tuning: Optimizing continuous prompts for generation.*

Liao, P., Li, X., Liu, X., & Keutzer, K. (2022). *The artbench dataset: Benchmarking generative models with artworks.*

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Martin, L., Ammanabrolu, P., Wang, X., Hancock, W., Singh, S., Harrison, B., & Riedl, M. (2018, April). Event representations for automated story generation with deep neural nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved from `http://dx.doi.org/10.1609/aaai.v32i1.11430` doi: 10.1609/aaai.v32i1.11430

Saleh, B., & Elgammal, A. (2015). *Large-scale classification of fine-art paintings: Learning the right metric on the right feature.*

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition.*

Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023). *Efficient streaming language models with attention sinks.*