Latent-Interp: A Deep Dive into Latent Spaces

Gannon Gonsiorowski

Introduction

In this project, I set out to explore and understand the structure of latent spaces generated by an autoencoder trained on the ATD12k dataset. The primary objective was to analyze how latent representations, which are derived from sequences of animated frames, encapsulate both visual content and motion information. Although the original plan included fine-tuning the autoencoder using a combination of KL divergence and reconstruction (MSE) losses, the fine-tuning process did not yield qualitatively distinct reconstructions. In fact, during training the finetuned VAE reached very low losses (KL = 0.00084 and MSE = 0.131), but a qualitative visual inspection revealed that the reconstructed images were all very similar, effectively averaging the input images rather than capturing dynamic details. This unexpected behavior suggested that the fine-tuning process was not contributing to a meaningful improvement in the latent representations, leading to a decision to rely on the pretrained VAE for further evaluation and latent space exploration. Ultimately, while the fine-tuning attempt provided useful insights into training metrics, it became clear that focusing on latent space analysis with the pretrained model was a more productive approach given the objectives of the project.

Background

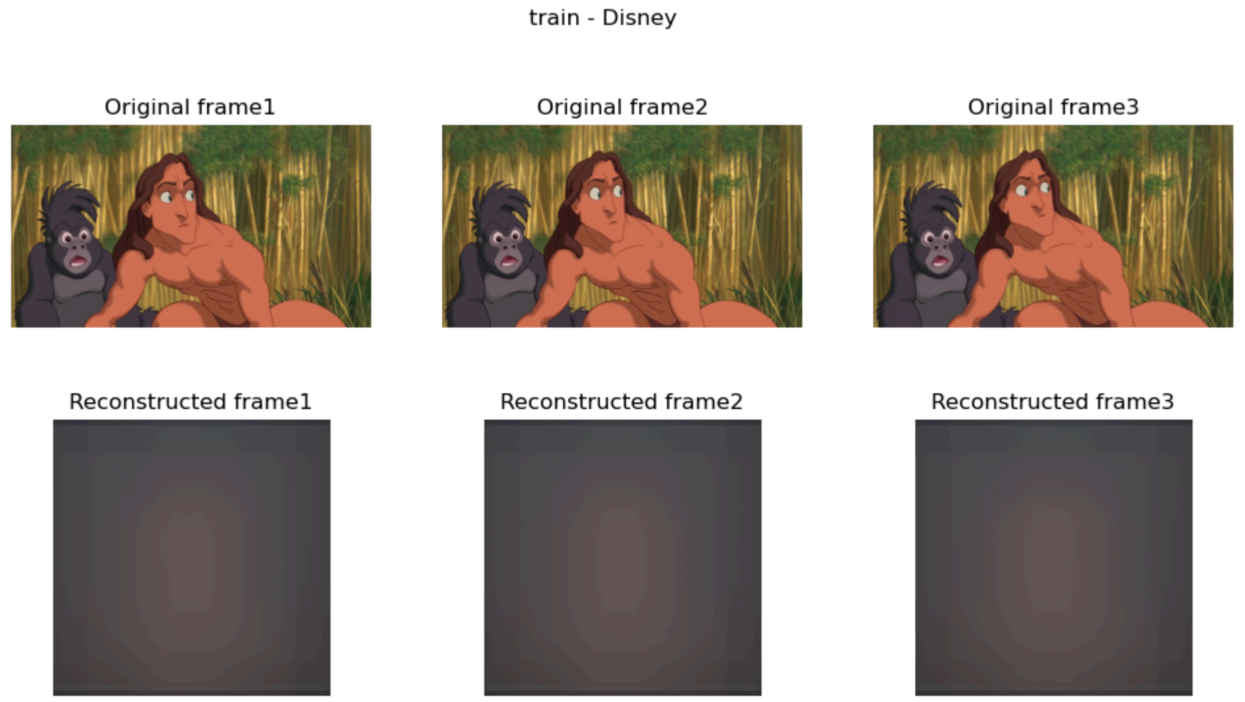
The project was inspired by my previous works on animated frame interpolation. The first time I worked on this project was 2 years ago, in which I trained a GAN to interpolate between frames 1 and 3 in image space. A year later I trained a diffusion model to do the same task with moderate success. Another year after that and my interests point towards autoencoders and latent spaces. Traditional autoencoders, especially variational autoencoders (VAEs), have been widely used to learn compressed representations of image data, while enforcing a continuous latent manifold through the use of KL divergence regularization. The literature (Kowalski et al. 2020, Gulrajani et al. 2016) shows that VAEs can effectively capture high-level features in data, and recent advances in visualization techniques such as UMAP and t-SNE have allowed researchers to gain further insights into the organization of latent spaces. The project proposal outlined a strategy to explore these transitions by both visualizing the latent trajectories and developing models that could predict intermediate latent representations. While my initial plans included extensive fine-tuning of the VAE, my experience with the autoencoder indicated that the pretrained model already provided a sufficiently rich latent space for the purpose of exploring interpolation and motion dynamics. This project therefore emphasizes latent space analysis over the fine-tuning process, with a detailed examination of both quantitative metrics and qualitative visualizations.

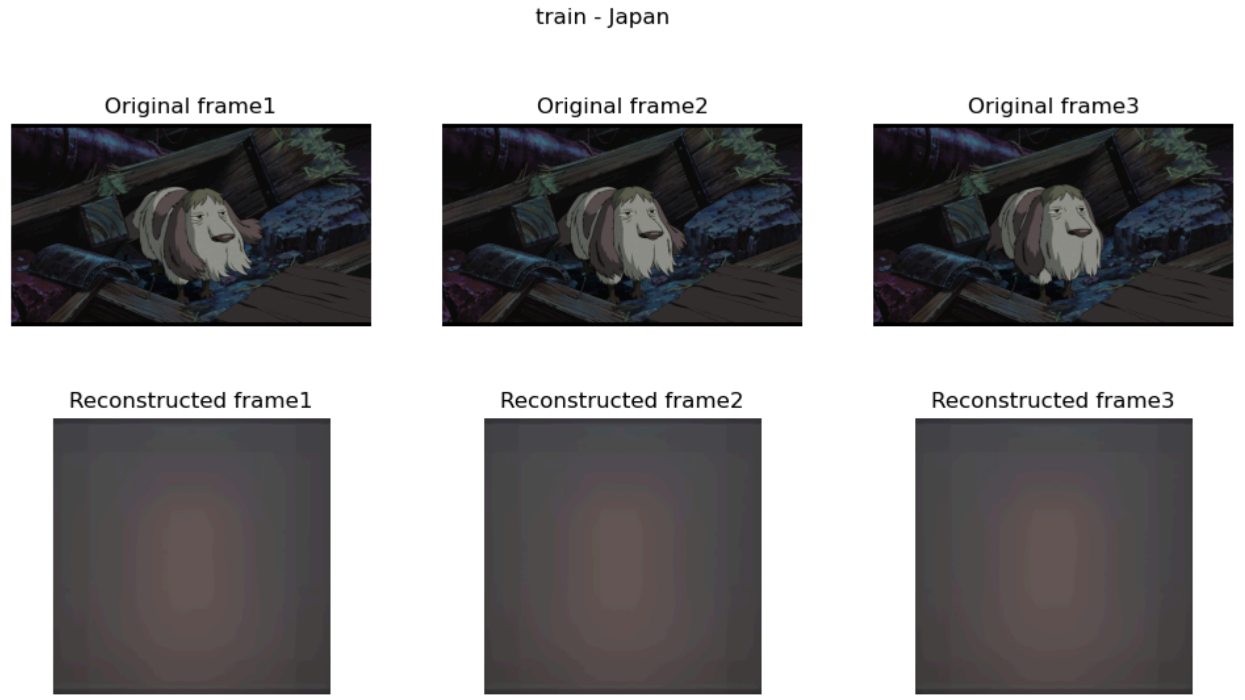
Data and Methods

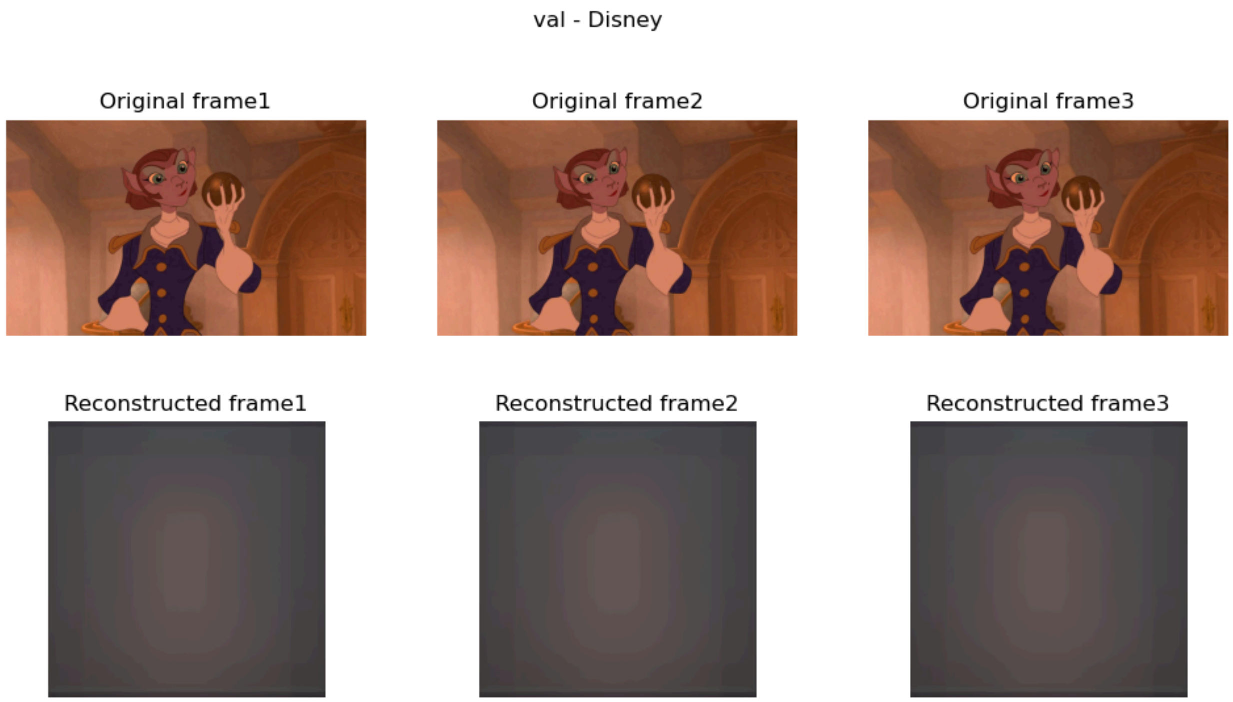
The data used in this project comes from the ATD12k dataset (Siyao et al. 2021), which consists of triplets of animated frames stored in well-structured directories for training, validation, and testing. Each subdirectory contains three latent files (latent1.npy, latent2.npy, latent3.npy) generated by an autoencoder, with each latent representation having the shape (4, 32, 32) and stored as float32 values. The autoencoder was originally designed to compress the visual and motion information from these frames into a latent space. The specific VAE I used was the kl-f8-anime2 model from hugging face. This model is used in many stable diffusion implementations because of its proficiency in handling animated content. This was an obvious choice considering my dataset is animated triplets from Disney and various anime films. My initial approach was to fine-tune this autoencoder using a combined loss function that included a reconstruction (MSE) term and a KL divergence term. The reconstruction loss is calculated as the mean squared error between the decoded image and the ground truth, ensuring that the output image closely resembles the input, while the KL divergence measures how far the latent distribution is from a prior (usually a standard Gaussian), encouraging a smooth and continuous latent space. Despite achieving promising numerical loss values during training, the decoded outputs remained nearly identical across epochs, suggesting that the model was converging to an average representation rather than capturing nuanced details. As a result, I shifted the focus of the project to utilize the pretrained VAE without fine-tuning for all subsequent analyses.

Evaluation and Results

Evaluation of the autoencoder and the subsequent latent space exploration was performed using both quantitative and qualitative methods. Initially, I fine-tuned the VAE on the ATD12k dataset. After 10 epochs, the model reached losses of KL=.00084 and MSE=.131. While this seemed great, I did a qualitative analysis before evaluating this model. The results were absolutely garbage. They show that the fine-tuned VAE likely learned to average all of the images it trained on. Earlier epochs with worse training metrics yielded the same results. Because of these results, I decided to skip the quantitative evaluation of this model. Below are the outputs of the fine-tuned model. The qualitative results are shown below.

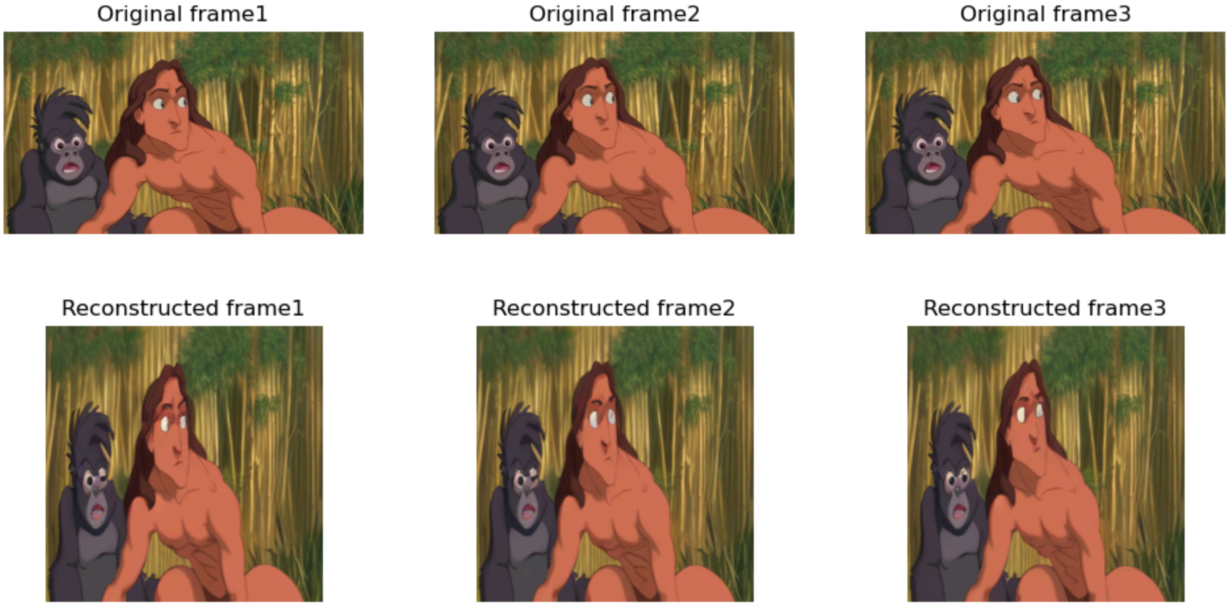






The key quantitative metrics included Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). MSE provides a straightforward measure of pixel-level reconstruction error, whereas PSNR, derived from MSE, offers a logarithmic scale that is more interpretable in terms of signal quality. SSIM, on the other hand, compares structural information between images, providing a more perceptually relevant assessment. When the pretrained VAE was evaluated on the entire dataset, it achieved an average MSE of 0.0009, an average PSNR of 32.0207, and an average SSIM of 0.9112. These values indicate that the autoencoder is capable of generating high-quality reconstructions that are both numerically and perceptually similar to the original images. The results of the qualitative analysis are below.

train - Disney



Original frame1

Original frame2

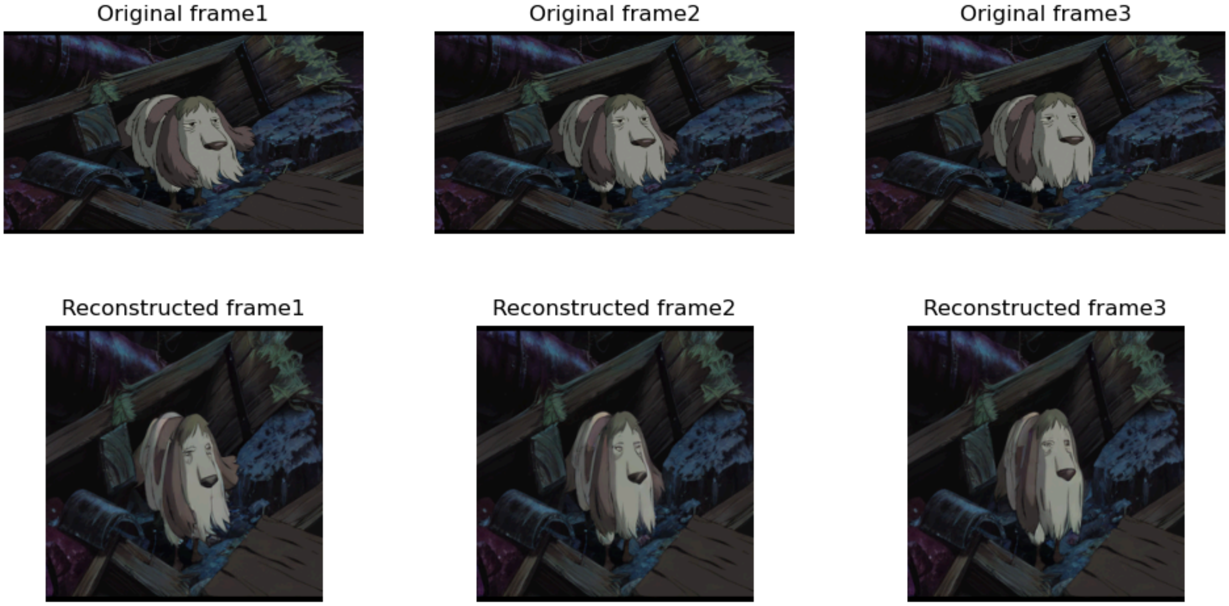
Original frame3

Reconstructed frame1

Reconstructed frame2

Reconstructed frame3

train - Japan



Original frame1

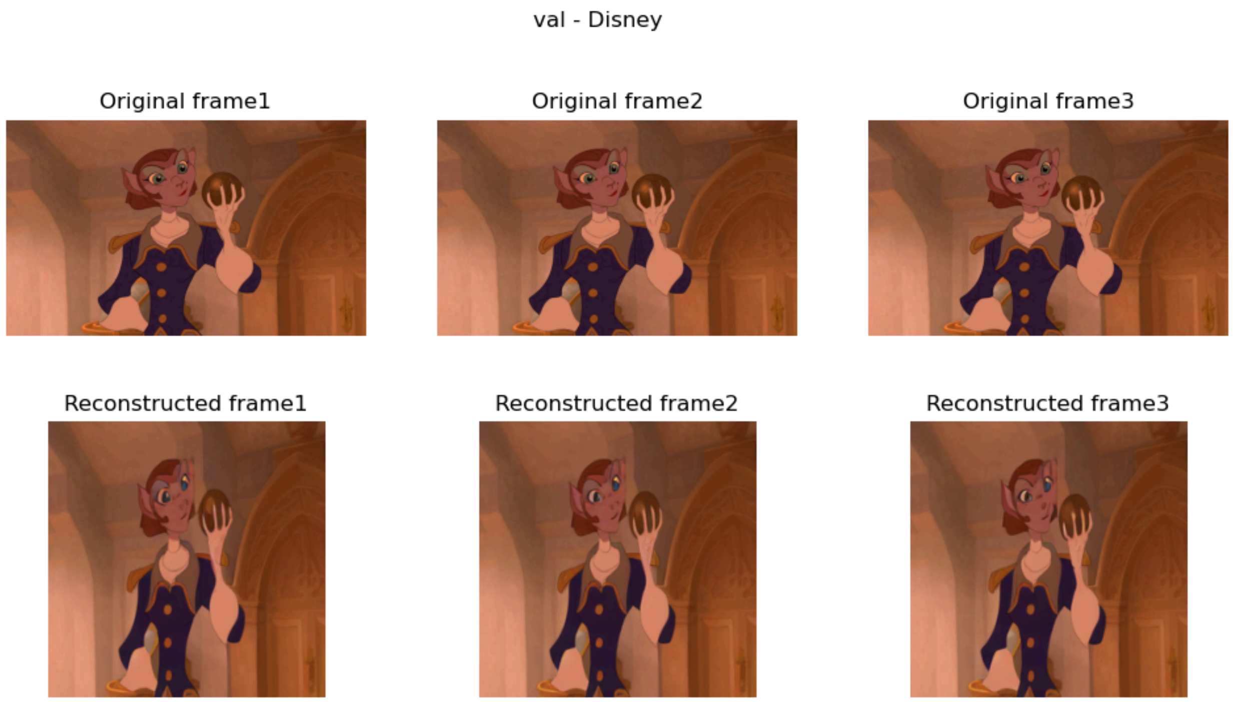
Original frame2

Original frame3

Reconstructed frame1

Reconstructed frame2

Reconstructed frame3



Additionally, various latent space exploration techniques were employed to assess the organization and continuity of the latent representations, thereby validating the effectiveness of the pretrained VAE for further analysis despite the failed fine-tuning attempts.

Latent Space Analysis

Due to time constraints, I was not able to put the latent space exploration into a good format for the report. Please see the latent_analysis.ipynb located on my Github repo for this project.

Outlook

Despite the initial challenges with fine-tuning the autoencoder, the project yielded significant insights into latent space behavior. The failure of the fine-tuning approach, where the reconstructed outputs converged to an average image, underscored the importance of relying on a well-established pretrained model when the primary focus is on latent space analysis rather than improving reconstruction quality. Moving forward, further research could investigate alternative fine-tuning strategies or additional regularization methods to refine the latent representations. Future work may also extend the analysis by incorporating more sophisticated interpolation techniques, additional quantitative metrics, and interactive visualization tools to explore temporal dynamics in greater depth. In retrospect, the project has demonstrated that a comprehensive analysis of latent spaces can be achieved through a combination of quantitative evaluations and diverse visualization techniques, and the insights gained here provide a strong foundation for subsequent work in generative modeling and animation synthesis. The lessons learned regarding the limitations of fine-tuning in this context, as well as the robustness of pretrained models, will inform future research directions, including potential applications in video frame interpolation and animation generation.

Overall, this project has met its objective of exploring latent spaces, providing a detailed characterization of how latent representations capture both spatial and motion information. The combination of quantitative metrics (MSE, PSNR, SSIM) and qualitative visual analysis has demonstrated that the pretrained VAE produces high-quality reconstructions. Additionally, the extensive latent space exploration using various visualization and interpolation methods offers valuable insights into the structure and dynamics of the latent space, paving the way for future work in deep representation learning and generative modeling.

Citations

Kowalski, M., Garbin, S. J., Estellers, V., Baltrušaitis, T., Johnson, M., & Shotton, J. (2020). Config: Controllable neural face image generation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16 (pp. 299-315). Springer International Publishing.

Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., & Courville, A. (2016). Pixelvae: A latent variable model for natural images. arXiv preprint arXiv:1611.05013.

Siyao, L., Zhao, S., Yu, W., Sun, W., Metaxas, D., Loy, C. C., & Liu, Z. (2021). Deep animation video interpolation in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6587-6595).