Procedural Content Generation Using Diffusion Models

SOUTHWESTERN UNIVERSITY

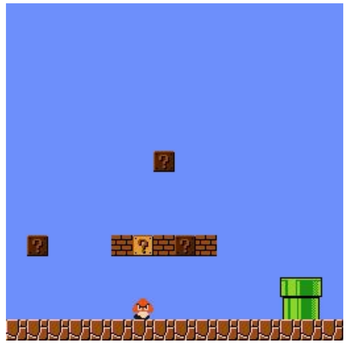
Olivia Kilday, Bess Hagan, Reid Williams, Emilio Salas

Introduction

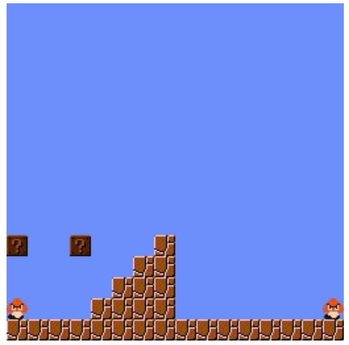
We present strategies to automatically assign descriptive captions to an existing level dataset, and train diffusion models [2] using both pretrained text encoders and simple transformer models trained from scratch. We assess the diversity and playability of the resulting levels, as well as their adherence to the given captions.

Diffusion Models

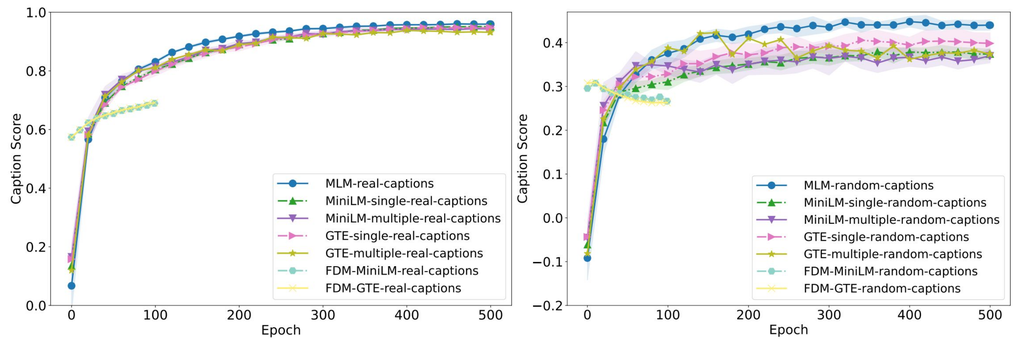
- Foundation of modern image/video generation
- Trained to remove noise from a given noisy image
- Eventually capable of generating content from noise
- Applied to one-hot encoded Mario levels
- Text embeddings of captions given for prompting
  - Both pretrained sentence embedding and custom masked language modeling transformer models



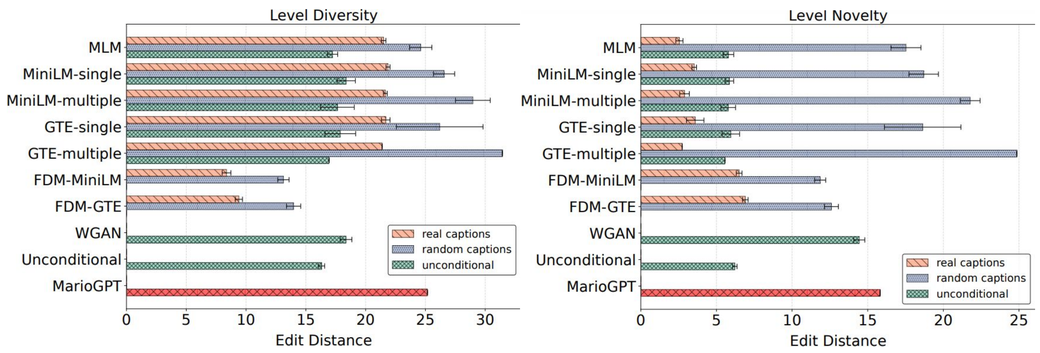
Real Level Scene: Caption:
“full floor. one enemy. a few question blocks. one platform. one pipe.”



Generated Scene: Input Prompt:
“floor with one gap. a few platforms. a few enemies. a few coins. one coin line. a few towers. one ascending staircase. a few question blocks.”
Actual Caption:
“full floor. two enemies. one ascending staircase. two question blocks.”
Caption Adherence Score: 0.478



Caption Score During Training: Shows how well generated levels match input prompts. Scores range from -1.0 to 1.0. Text-conditioned diffusion models, and Five-Dollar-Models (FDM) [1] are compared. FDM training stops early to avoid overfitting. Left: high scores for actual game captions. Right: lower scores for random captions. Diffusion with our MLM transformer performs best, over FDM and diffusion with pretrained MiniLM [5] and GTE [6] sentence embedding models.



Left: Level Diversity: How different levels are from other generated levels. Right: Level Novelty: How different levels are from the original Mario levels. Models generate levels in three ways: from real captions, from random captions, and unconditionally (no caption). Random captions create more diverse and novel levels. Real captions are less novel because they match the training data. Real captions create more diverse levels than unconditional generation. We compare against the Five-Dollar-Model (FDM) [1], WGAN models [3], unconditional diffusion, and MarioGPT [4].

Results

- Text conditional diffusion models are far better in caption score than FDM, despite FDM’s quicker training time
- Despite its longer training time, GTE models were generally no better than other models
- The novelty of diffusion models is easily controllable from the caption
  - Real captions, those similar to existing levels, lead to output similar to real levels
  - Random captions, very different from training data, leads to very diverse and novel levels
- All models struggled with random captions, seeing a significant drop in caption adherence score
- Our MLM model, the smallest text encoder here, still consistently outperformed all other models
- MarioGPT is a special case, performing well, but on a very limited captioning system

References

[1] Merino, T.; Negri, R.; Rajesh, D.; Charity, M.; and Togelius, J. 2023. The Five-Dollar Model: Generating Game Maps and Sprites From Sentence Embeddings. In Artificial Intelligence and Interactive Digital Entertainment.
[2] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In Computer Vision and Pattern Recognition. IEEE.
[3] Volz, V.; Schrum, J.; Liu, J.; Lucas, S. M.; Smith, A. M.; and Risi, S. 2018. Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network. In GECCO. ACM.
[4] Sudhakaran, S.; Gonzalez-Duque, M.; Freiberger, M.; Glanois, C.; Najarro, E.; and Risi, S. 2023. MarioGPT: Open-Ended Text2Level Generation Through Large Language Models. In NIPS.
[5] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1
[6] https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5