SceneGenie: Scene Graph to Image via CLIP Embeddings and Diffusion Model based Generation

General Info

Project Title: SceneGenie: Scene Graph to Image via CLIP Embeddings and Diffusion Model based Generation

Contact Person: Azade Farshad, Yousef Yeganeh

Contact Email: azade.farshad@tum.de, y.yeganeh@tum.de

Project Abstract

The goal of this project is to perform image generation from scene graphs[1]. There are two objectives in this project to improve the image generation quality: 1) Using the state-of-the-art CLIP[6] or data2vec[7] text to image embeddings for scene graph feature extraction, 2) Deploying diffusion models[4,5] into the SG2Im framework.

Background and Motivation

Image generation is an important task and has gained a lot of attention in recent years. Scene graphs are useful tools for easier manipulation[2] of the scenes and they have been recently used in modelling surgery rooms for action recognition. Learning a good feature representation and having a strong generator network are important aspects of image manipulation and generation. In this project, we focus on two aspects of representation learning for image generation inspired by the recent DALLE2[8] framework:
1) Contrastive text to image embeddings for scene graph feature extraction, 2) Diffusion models for image generation.

Technical Prerequisites

- Good background in statistics
- Good background in machine learning, deep learning
- Good skills in Python
- Good skills in PyTorch

Benefits:

- Weekly supervision and discussions
- Possible novelty of the research
- Possible publication

Students' Tasks Description

Students' tasks would be the following:

Groups 1 & 2:

- Understanding the underlying methods
- Evaluation on Visual Genome + COCO / Clinical dataset[3]

- Testing and documentation.

Groups 1,2:

- Familiarize with SG2Im framework
- Deploy CLIP embeddings in SG2Im framework
- Familiarize with diffusion models
- Deploy diffusion models in SG2Im framework

- **Work-packages and Time-plan:**

	Description	#Students	From	To
WP1	Familiarizing with the literature.	4	05.05	12.05
WP2	Familiarizing with the required frameworks. Come up with a detailed time-plan (gantt)	4	12.05	19.05
WP3	Employing CLIP embeddings into SG2Im	2	19.05	02.06
WP4	Adapting the generator to diffusion models	2	19.05	02.06
WP5	Evaluation of the implemented method	4	02.06	16.06
WP6	Comparison to related work + Preparing midterm presentation	4	16.06	23.06
M1	Intermediate Presentation II	4	23.06	
WP7	Familiarizing with clinical data, data pre-processing	4	23.06	01.07
WP8	Implement and Evaluate WP3/WP4 & WP6 on medical data	4	01.07	14.07
WP9	Testing and Documentation	4	14.07	28.07
M2	Final Presentation	4	28.07	

References

1. Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1219-1228).
2. Dhamo, H., Farshad, A., Laina, I., Navab, N., Hager, G. D., Tombari, F., & Rupprecht, C. (2020). Semantic image manipulation using scene graphs. In CVPR.
3. Özsoy, E., Örnek, E. P., Eck, U., Tombari, F., & Navab, N. (2021). Multimodal Semantic Scene Graphs for Holistic Modeling of Surgical Procedures. arXiv preprint arXiv:2106.15309.
4. Nichol, A. Q., & Dhariwal, P. (2021, July). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (pp. 8162-8171). PMLR.
5. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34.

6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.
7. Baevski, A., Hsu, W. N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.