Spring 2023 Boston University

**CS 585: Image and Video Computing**

**Final Project Topic Proposal**

Wonyl Choi U29967430, Duc Minh Nguyen U17531382, Quan Pham U52532940

# Description of the Final Project Topic

## **Topic title**

## Text-to-image generation using Hilbert-Schmidt Independence Criterion

## **Objective**

The main goal of the project is to investigate self-supervised representation learning in text-to-image generation task. More specifically, the main objectives of the projects are as follows:

* Understand and build the generative adversarial network (GAN) based model for the text-to-image generation task
* Study the efficacy of the contrastive learning framework on the text-to-image generation task and implement the Hilbert-Schmidt independence criterion (HSIC) loss for the task
* Measure the performance of the model and perform the ablation study for each design choice made.

## **Summary**

The text-to-image generation task aims to generate an image from a representation vector sampled from the manifold conditioned by the embedding of a given textual description. Naturally, GAN has been a popular choice for the backbone of models tackling the task, more specifically, to model translation from the representation space to the image space.

Particularly on the task of text-to-image generation, there have been several recent publications of reaching state-of-the-art or near state-of-the-art performance by utilizing GAN with contrastive learning. For example, Lafite[[1]](#footnote-0) used pre-trained CLIP[[2]](#footnote-1) for the alignment of the image and text embeddings and also used pseudo-text embeddings generated from the real image for text-free training enhancing the performance. XMC-GAN[[3]](#footnote-2) used mutual information maximization for contrastive discriminator loss. The performance improvements at the incorporation of the contrastive learning on the previously suggested GAN-based models also have been investi- gated (Ye et al. (2021)[[4]](#footnote-3))

Building upon the recent advances, we aim to investigate the effect of contrastive learning on the text-image generation task. In particular, we aim to build a model using the Hilbert-Schmidt Independence Criterion (HSIC)[[5]](#footnote-4). The HSIC is originally suggested as an independence measure of two probability distributions in the reproducing kernel Hilbert space (RKHS). The attractiveness of HSIC arises from its theoretical aspect. In contrast to the methods equipping indirect mutual information estimators, HSIC allows to optimize statistical dependencies directly. Here, the latent space subjected to the optimization can be interpreted as the probability manifold.

The efficacy of HSIC on self-supervised representation learning on image classification tasks has been studied by Li et al. (2021)[[6]](#footnote-5), but to our best knowledge, there is no literature investigating the application of HSIC-based contrastive loss on text-image generation.

## **Background**

Contrastive learning is a general framework of learning representation space of the given data. The main idea of contrastive learning is that the metric or the distance measure must correspond to the similarity of data points. Alternative to the metric learning view, it is possible to connect the framework to the idea of mutual information maximization. By maximizing the estimated mutual information (MI) of different views of the input data, unsupervised or self-supervised representation learning that ensures more generalizability can be achieved. The difficulty of estimating MI can be mitigated by using the Hilbert-Schmidt Independence Criterion (HSIC).

In the vision community, the effectiveness of the contrastive learning framework on the self-supervised visual representation learning has been empirically demonstrated, namely by simCLR, Barlow Twins, and BYOL. While the application of learnt representation is usually found in discriminative tasks such as classification or clustering, it has been shown that the contrastive learning framework can be incorporated with the generative models, for example, infoGAN.

# Related Papers

We plan to establish the generative model for text-to-image generation following Lafite1**.** The implementation of the model can be found in Github[[7]](#footnote-6). We plan to modify the model especially to use HSIC as the contrastive learning loss. The HSIC loss implementation will be done by ourselves, following the original theoretical paper5, taking the JAX implementation made by Li et al. (2021)6 as the main reference.

For evaluation, we plan to report Fréchet inception distance (FID)[[8]](#footnote-7) values.

Other extensions of the model will be made if time permits.

# Datasets

One of the datasets that are commonly used in the literature of text-to-image generation is MS-COCO[[9]](#footnote-8). The dataset contains 5 captions per image and includes iconic objects, iconic scenes and non-iconic images. The statistics of the dataset and sample data are presented below.

| **Dataset** | **Train** | **Validation** | **Caption per Image** |
| --- | --- | --- | --- |
| MS-COCO | 82k | 40k | 5 |
| Visual Genome | 108k | 5k | 3.3(Avg) |
| Flickr30k | 20k | 5k | 5 |

Table. Statistics of Datasets

****

Fig. Sample images and captions of MS-COCO

# Contributions

W.C. conceived of the project idea. W.C. will manage the basic model structure and implement HSIC loss. Q.P. will perform the experiment reproducing the result of Zhou et al. (2022)1 and the result visualisation module. D.N. will be responsible for data curation, collection, and implementation of the data input and evaluation module. W.C. and D.N. / Q.P. will be responsible for designing and implementing the training strategies. All project members will discuss the results and contribute to the final manuscript.

1. Zhou, Yufan, et al. "Towards language-free training for text-to-image generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. [↑](#footnote-ref-0)
2. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021. [↑](#footnote-ref-1)
3. Zhang, Han, et al. "Cross-modal contrastive learning for text-to-image generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021. [↑](#footnote-ref-2)
4. Ye, Hui, et al. "Improving text-to-image synthesis using contrastive learning." *arXiv preprint arXiv:2107.02423* (2021). [↑](#footnote-ref-3)
5. Gretton, Arthur, et al. "A kernel statistical test of independence." *Advances in neural information processing systems* 20 (2007). [↑](#footnote-ref-4)
6. Li, Yazhe, et al. "Self-supervised learning with kernel dependence maximization." *Advances in Neural Information Processing Systems* 34 (2021): 15543-15556. [↑](#footnote-ref-5)
7. https://github.com/drboog/Lafite [↑](#footnote-ref-6)
8. Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems* 30 (2017). [↑](#footnote-ref-7)
9. Lin, Tsung-Yi, et al. "Microsoft COCO: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014. [↑](#footnote-ref-8)