Создание генератора изображений в ограниченных условиях.

mini-CREAM

Абстракт. В последнее время генераторы изображений на основе нейронных сетей получили большую популярность. Качество изображений по сравнению с первыми моделями dall-e[1], stable diffusion[2]. Однако, как правило, чтобы создать работающий генератор требуется много данных, и вычислительных мощностей. Наш эксперимент показывает и доказывает, что можно собрать работающий генератор изображений в ограниченных условиях.

Обучение производилось на одном графическом процессоре T4[3] 30 минут. Код и веса находятся по ссылке: https://github.com/companys1234/mini-CREAM

Введение. В основном современные модели обучаются на больших серверах GPU. беспрерывно несколько месяцев. Это оставляет большой отпечаток как на экологии, так и на финансовых сбережениях компании. До этого были работы посвященной этой теме[9]. Как правило, эти решения были созданы при неограниченных возможностях. Мы представляем собственную модель, созданную в ограниченных условиях, которая при дальнейшем обучении и улучшение может минимизировать проблемы больших затрат на обучение, и инференс модели.

Предыдущие работы.

Diffusion models[4] - это модели генерирущие материал с использованием диффузионного процесса. Диффузионный процесс - это процесс генерации материала, путём прямого процесса (добавления шума к изображению), и обратного

процесса (удаление шума из изображений). За t шагов.

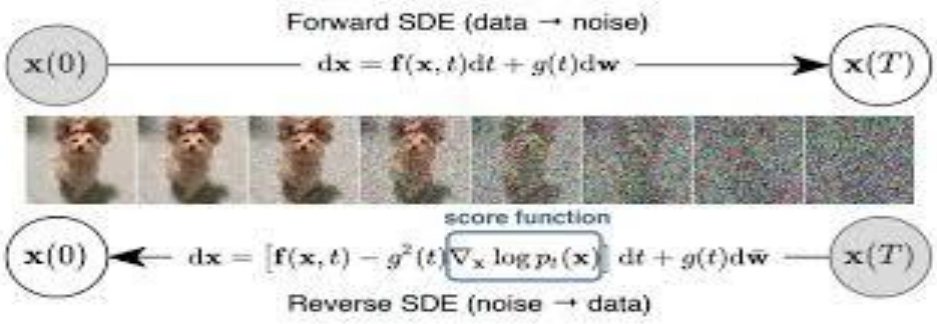

Рис 1. Процесс работы диффузионных моделей

Stable diffusion - это диффузионная модель генерации изображения по текстовой подсказке, с использованием cross-attention. Диффузионный процесс модели основан на архитектуре Unet[13].

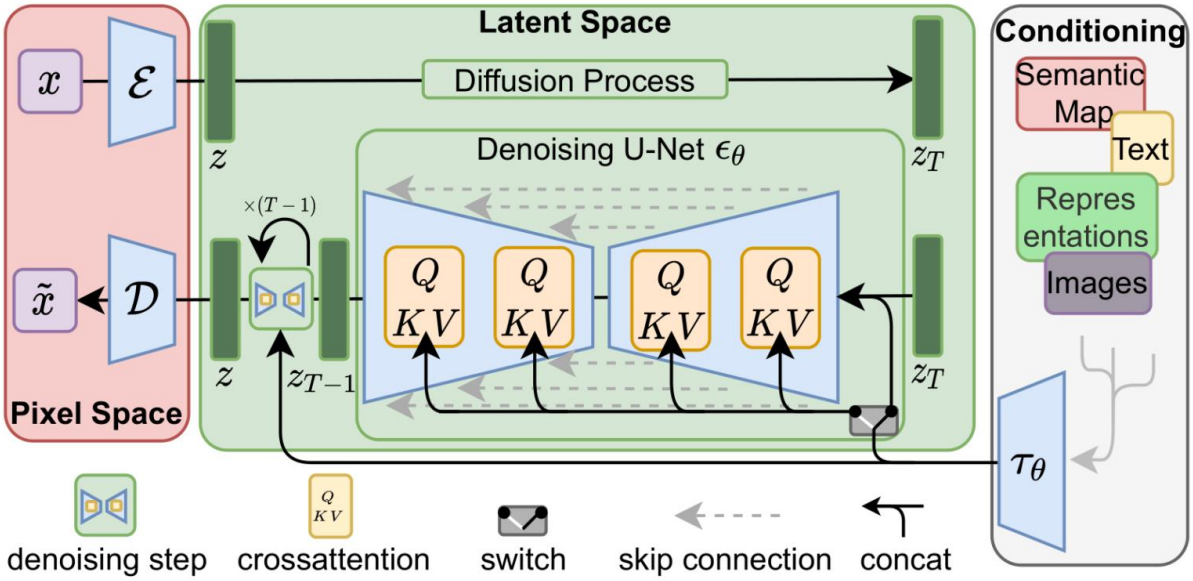

Рис. 2.Архитектура Stable diffusion

Cross-attention - это один из видов внимания, введёный в статье[5]. Этот вид внимания позволяет декодировщику сосредоточиться на наиболее важных деталях последовательности полученной от кодировщика

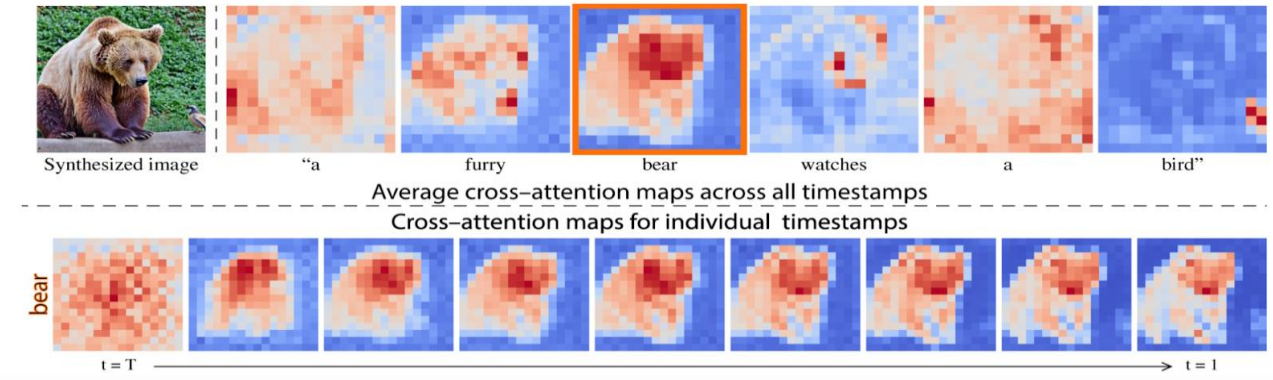

Рис 3.cross-attention в stable diffusion

Clip[6] -это эмбединг связывающий изображения, и подписи к ним. Clip - это мощный инструмент, который связывает изображения и язык, что может помочь в разработке разнообразных решений.

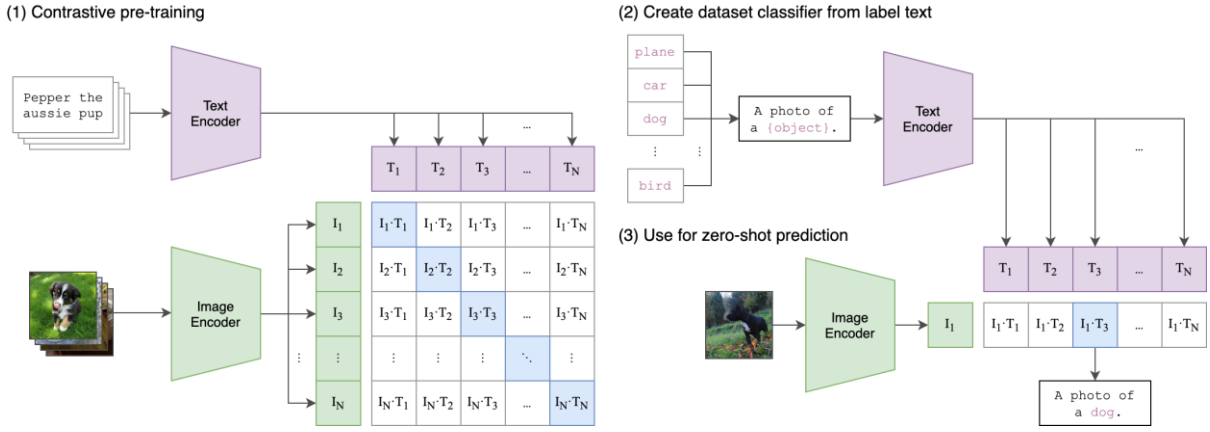

Рис 4.Архитектура Clip.

MOE(mixture-of-experts)[7] - это архитектура нейронной сети, которая основана на том,что мы обучаем сеть по отдельным частям(экспертам) в зависимости от входных данных. Инференс модели происходит тем же образом, активируется не вся сеть, а лишь часть. Выбор экспертов

проводится через отдельную сеть(Router).

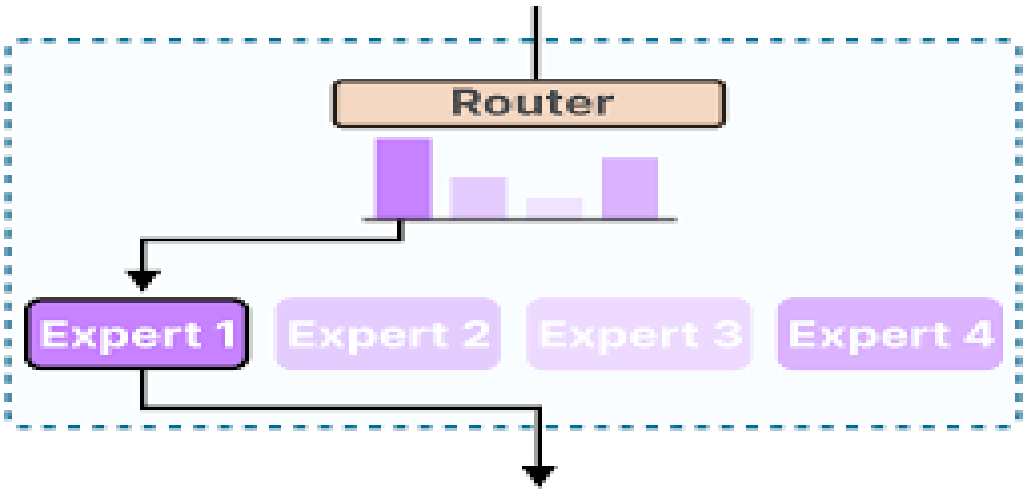

Рис 5. Архитектура MOE

Архитектура: архитектура mini-cream во многом основана на архитектуре stable diffusion. Однако, mini-cream во многом самобытен, и дополняет stable diffusion.Основные нововведения сделанные в mini-cream:1 language filter, 2 MOE block, 3 оценщик изображений.

Language filter - это языковая модель(например:Deepseek-R1[8]) которая по заданному контексту дополняет подсказки пользователей, а после подает эти подсказки остальной части сети.

MOE block - это часть сети которая находиться в середине и основана на архитектуре mixture-of-experts. Прошлые работы[9] показали то, что архитектуры с MOE намного эффективнее, чем архитектуры без MOE.

Оценщик изображений - оценивает сходство нескольких изображений с текстовой подсказкой, и выбирает наилучшее. Наш оценщик использует как метрику FID. При желании, оценщик может быть заменён на человеческие оценщики[10].

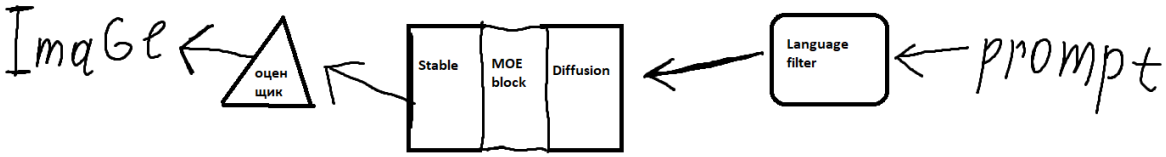

Рис 6. Архитектура mini-CREAM

Результаты: эксперименты проводились с нашей моделью, kandinsky[11], и midjourney[12]. Метрики которые были использованы: FID[14], IS[15], SSIM[16]. Промпты на которых проверяли: orange, apple.

	kandinsky	midjourney	Mini-CREAM
FID	459,27	171,27	444,60
IS	1,08	1,02	1,05
SSIM	0,3694	0,2319	1,000

Табл 1. Результаты проведённых экспериментов.

Заключение и вывод. В конечном итоге, несмотря на то, что mini-CREAM не превзошёл популярные модели в одинаковых условиях, во многом ему удалось сравнится в метриках. Возможно, в дальнейшем mini-CREAM сможет превзойти эти модели путём увеличения параметров самой

модели, увеличения данных, и увеличения мошьностей для обучения. Ознакомиться с примером работы mini-CREAM вы можете в приложении А.

Ссылки и ресурсы.

[1] - Ramesh A. et al. Zero-shot text-to-image generation //International conference on machine learning. – Pmlr, 2021. – С. 8821-883

[2] - Rombach R. et al. High-resolution image synthesis with latent diffusion models //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2022. – С. 10684-10695.

[3] - Datasheet_NVIDIA_T4_Virtualization

[4] - Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851.

[5] - Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.

[6] - Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning. – PmLR, 2021. – С. 8748-8763.

[7] - Cai W. et al. A survey on mixture of experts in large language models //IEEE Transactions on Knowledge and Data Engineering. – 2025.

[8] - Guo D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning //arXiv preprint arXiv:2501.12948. – 2025.

[9] - Xue Z. et al. Raphael: Text-to-image generation via large mixture of diffusion paths //Advances in Neural Information Processing Systems. – 2023. – Т. 36. – С. 41693-41706.

[10] - Ouyang L. et al. Training language models to follow instructions with human feedback //Advances in neural information processing systems. – 2022. – Т. 35. – С. 27730-27744.

[11] - Arkhipkin V. et al. Kandinsky 3.0 technical report //arXiv preprint arXiv:2312.03511. – 2023.
[12] - https://www.midjourney.com/
[13] - Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation //Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. – Springer international publishing, 2015. – С. 234-241.
[14] - Heusel M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium //Advances in neural information processing systems. – 2017. – Т. 30.
[15] - Barratt S., Sharma R. A note on the inception score //arXiv preprint arXiv:1801.01973. – 2018.
[16] - Nilsson J., Akenine-Möller T. Understanding ssim //arXiv preprint arXiv:2006.13846. – 2020.

Приложение А. Пример работы mini-CREAM.

Рис 5. Изображения созданные mini-CREAM по запросу dog.