Введение

creme_brulee

В данном отчете описывается проделанная работа для обучения моделей генерации изображений. Лучшая модель смогла достичь FID 0.35. Выполненная работа доступна по https://github.com/Professor322/genai_task

Описание задания

В поставленном задании требовалось обучить модель генерировать изображения из обработанного набора данных Food101, где разрешение каждой картинки составляет 64 на 64 пикселя.

Выбор базовой модели

State-of-the-art подходами в области генерации изображений являются диффузионные модели. В то же время, их недостатком является скорость работы на этапе применения, но поскольку в условии задачи не было введено дополнительных ограничений на время работы выбранного подхода, мною было принято решение остановиться на модели из семейства диффузионных. Более конкретно, в качестве базовой модели была выбрана improved diffusion от коллег из openAI. Я не стал использовать подходы на основе сжатия исходного изображения в латентное пространство, так как данные подходы предполагают работу с изображениями высокого разрешения, а мы работаем с картинками размера 64x64.

Основными улучшениями в improved diffusion являются косинусоидный режим добавления шума к изображению, так как он более плавно разрушает отличительные черты изображения, а также альтернативная функция потерь на основе variational lower bound (VLB) и MSE, благодаря которой FID получается лучше. На основе VLB + MSE авторы статьи предложили способ, ускоряющий работу модели на этапе применения, но эта техника не была применена в данной работе.

Эксперименты

В качестве экспериментов было решено обучить 2 диффузионные модели: одну со среднеквадратичной ошибкой в качестве функции потерь, вторую - на основе гибридной функции потерь, предложенной в статье “improved diffusion”. Предполагается, что модель, оптимизирующая VSB + MSE, достигнет лучших результатов.

Расписание добавление шума будет иметь косинусоидный вид для обеих моделей. В литературе, связанной с диффузионными моделями, чаще всего в качестве количества шагов добавления шума, авторы выбирают 1000 и более. Мною была сделано предположение из-за небольшой размерности входных изображений, может потребоваться меньше шагов зашумления. Таким образом, модель на каждой итерации предсказания шума, по сути, будет предсказывать более интенсивный шум.

Обе модели будут условными, так как это позволит контролировать класс генерации изображений. В первую очередь, это необходимо для того, чтобы количество сгенерированных изображений для каждого класса не отличалось от количества настоящих изображений для подсчета FID.

Детали подсчета FID

FID (Frechet Inception Distance) был посчитан на основе 64 весов из Inception V3 модели.

Результаты

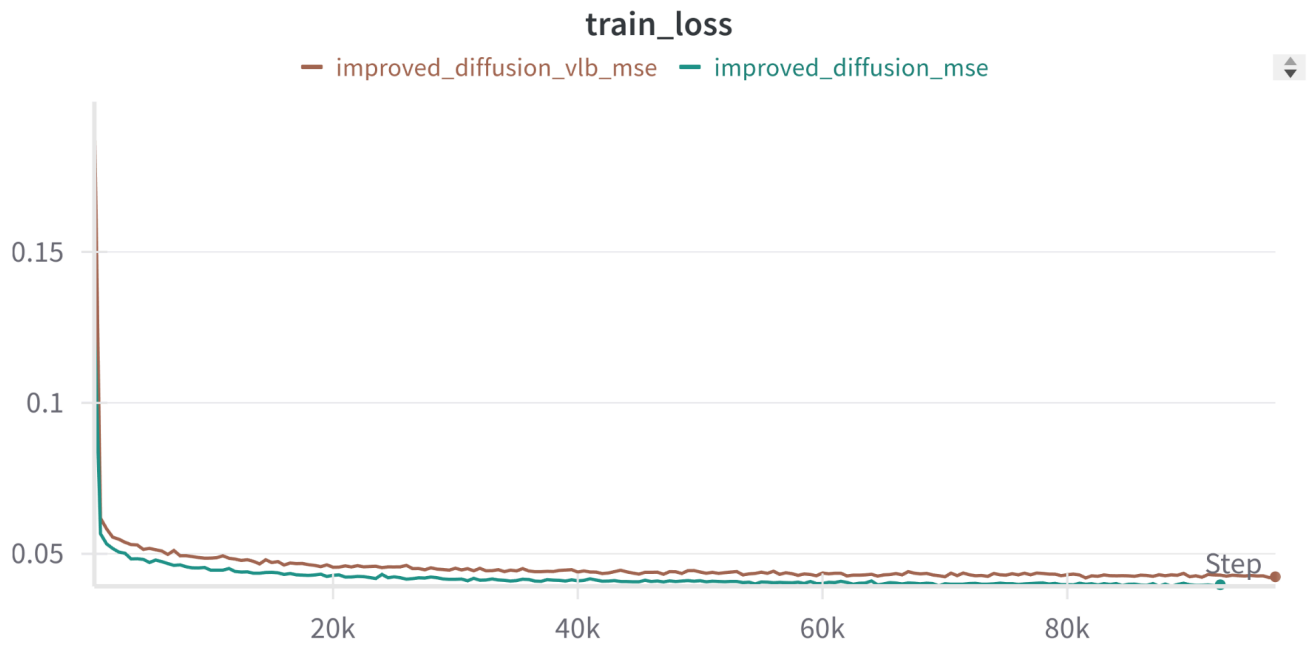

Рис 1. Изображение иллюстрирует изменение значения функции потерь на каждой итерации. Значение функции потерь для VLB + MSE выше, так как слагаемое VLB используется в качестве регуляризатора.

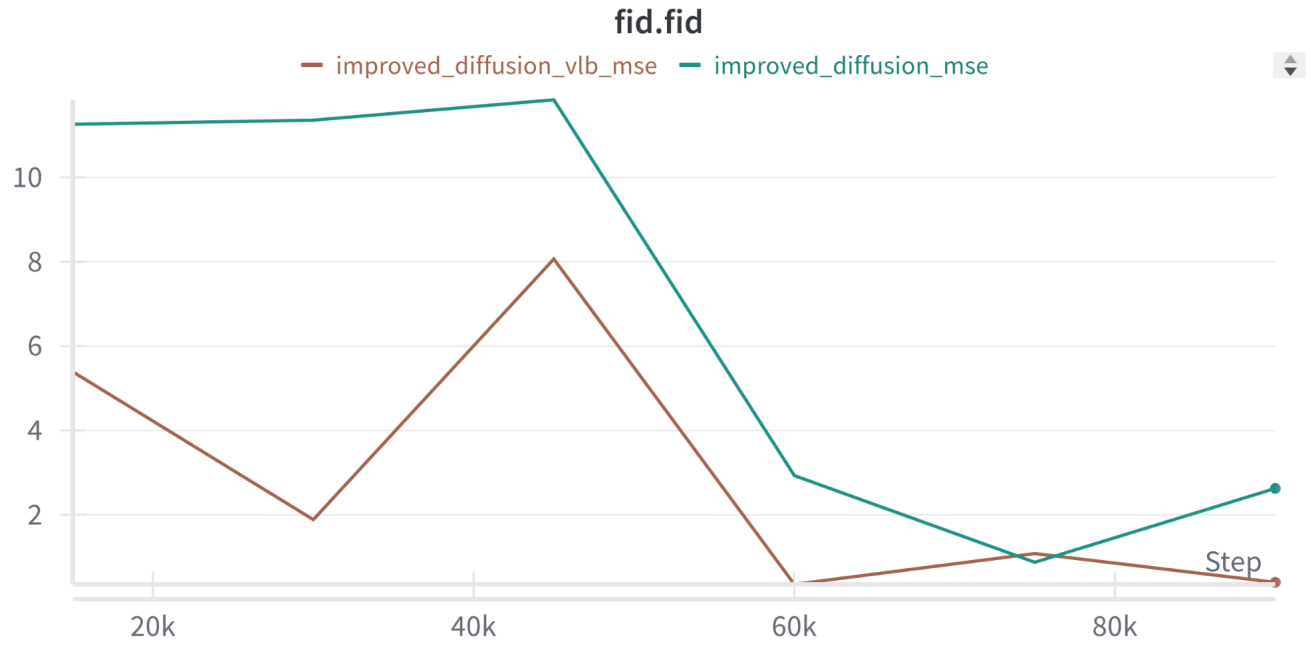

Рис 2. Практически на протяжении всего обучения FID модели VLB + MSE был ниже чем у MSE модели.

Несмотря на то, что значение функции потерь для VLB + MSE было постоянно выше, качество генерации у VLB + MSE модели получилось лучше (FID ниже). Таким образом гипотеза подтвердилась и, действительно, оптимизируя гибридную функцию потерь, мы можем достичь модель лучше по качеству на основе FID.

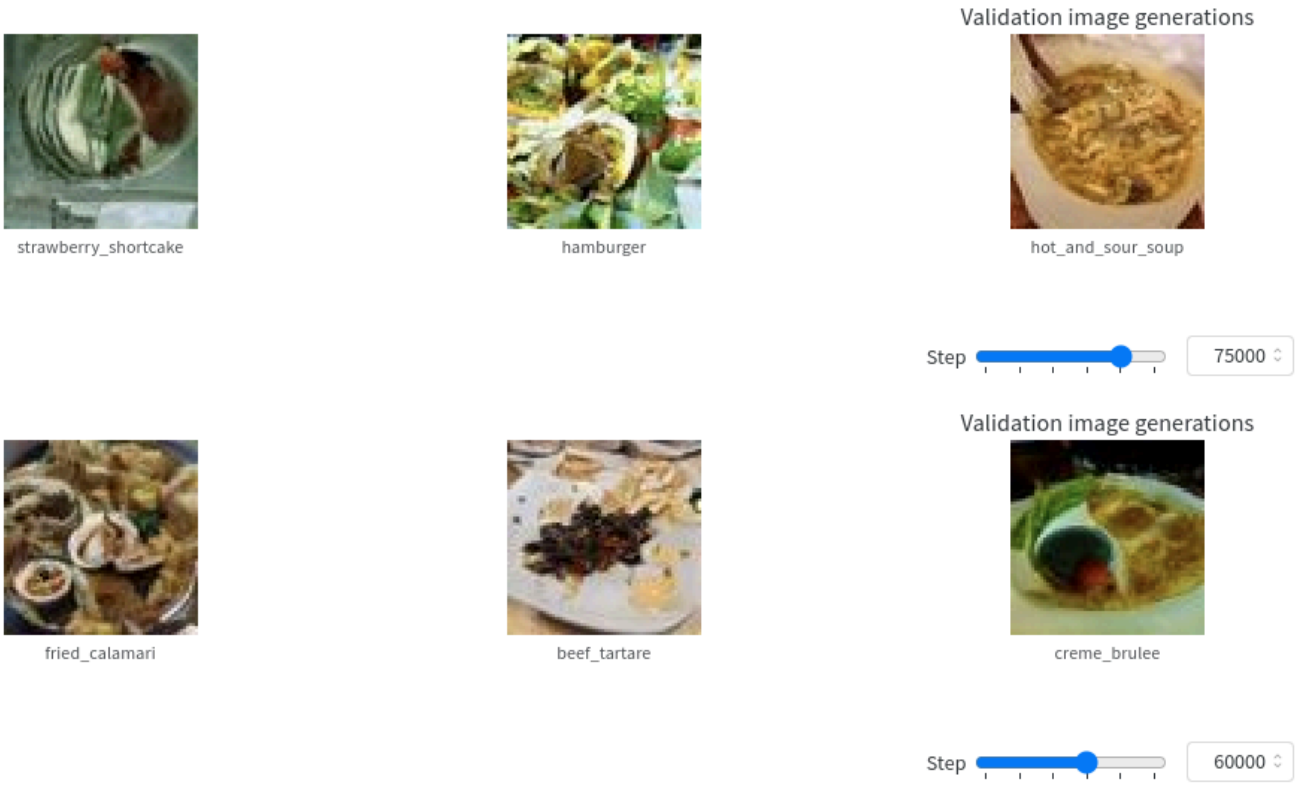

Рис 3. Изображения сгенерированные лучшими моделями. Верхний ряд MSE 75000 итераций, FID 0.87. Нижний ряд VLB + MSE 60000 итераций, FID 0.35.

Стоит также обратить внимание на то, что минимальный FID получился за меньшее количество итераций обучения у VLB + MSE.

Больше генераций

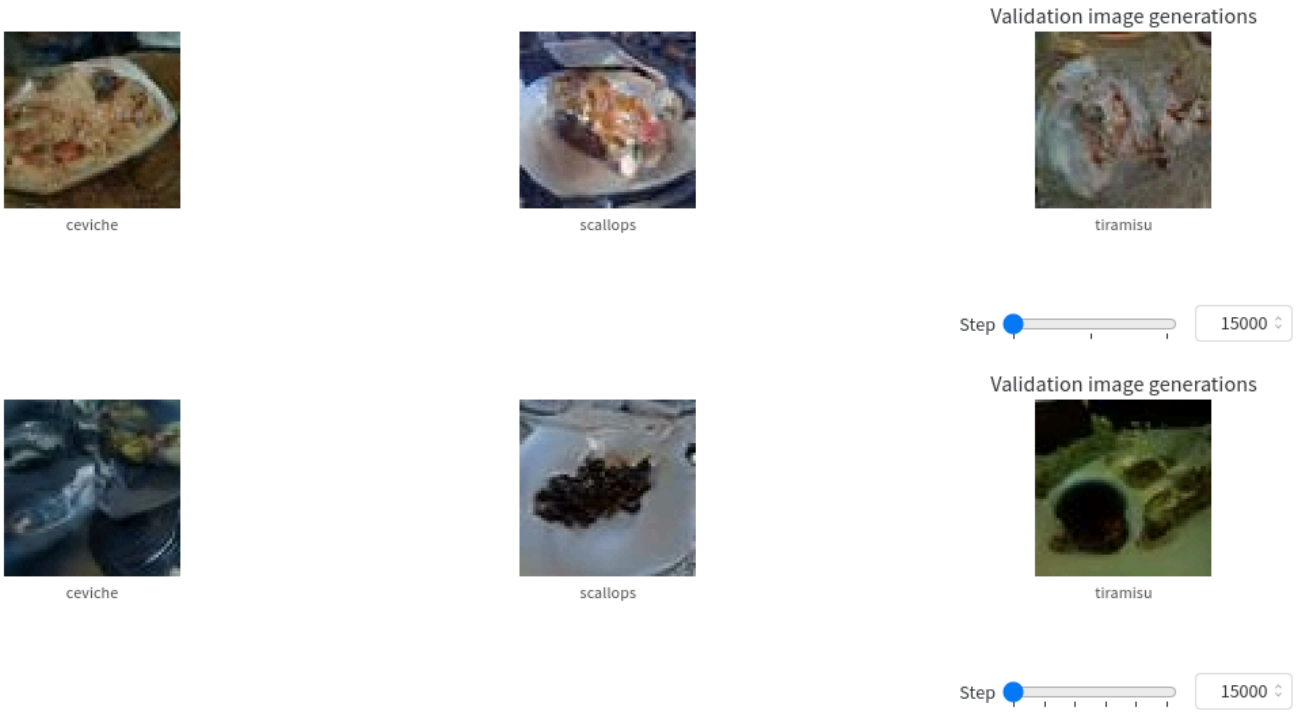

Рис 4. Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 15к итераций

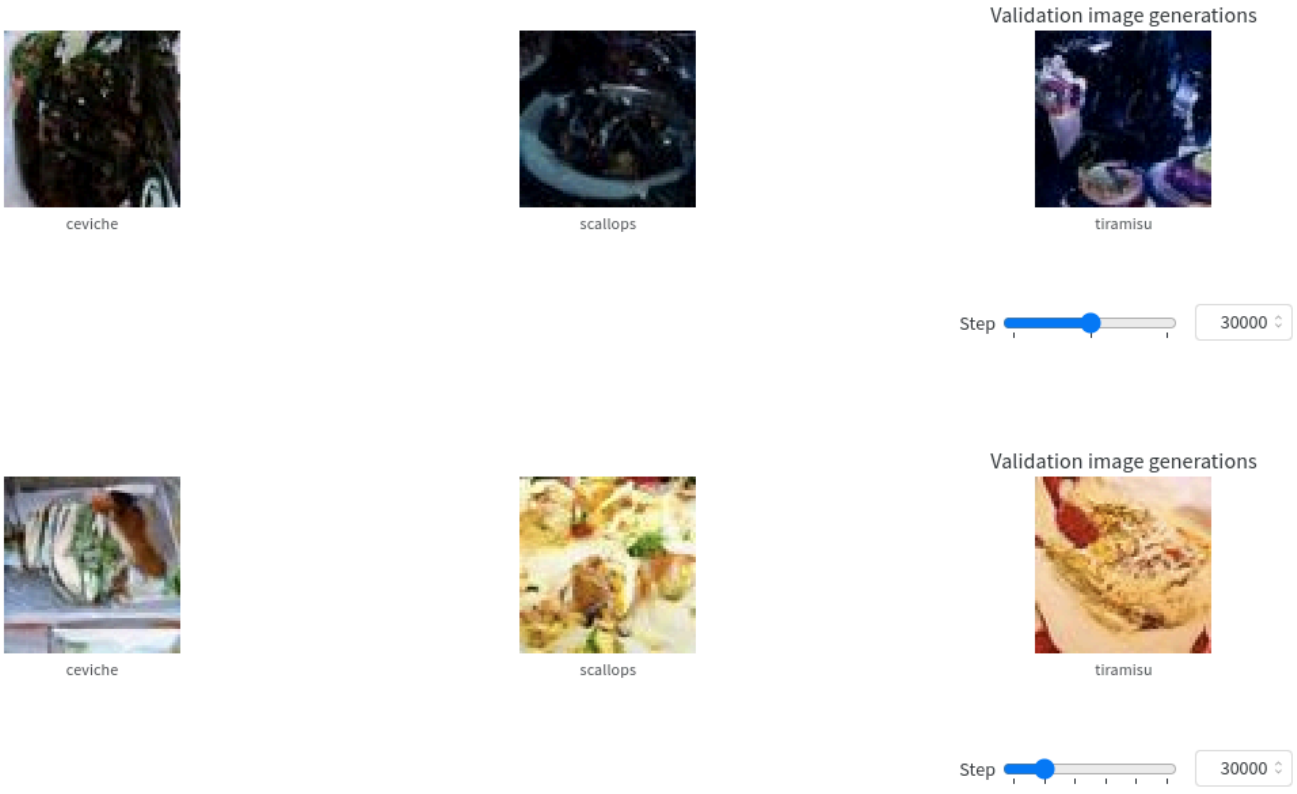

Рис 5.Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 30к итераций

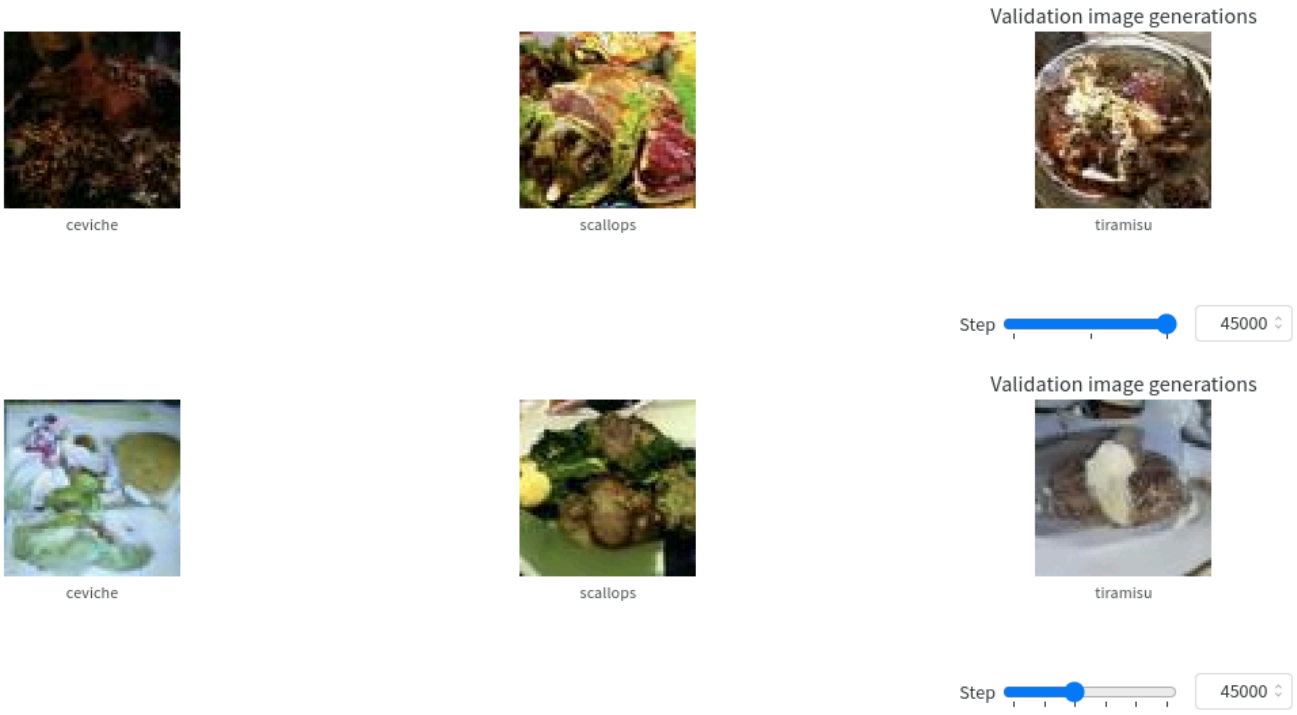

Рис 6. Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 45к итераций

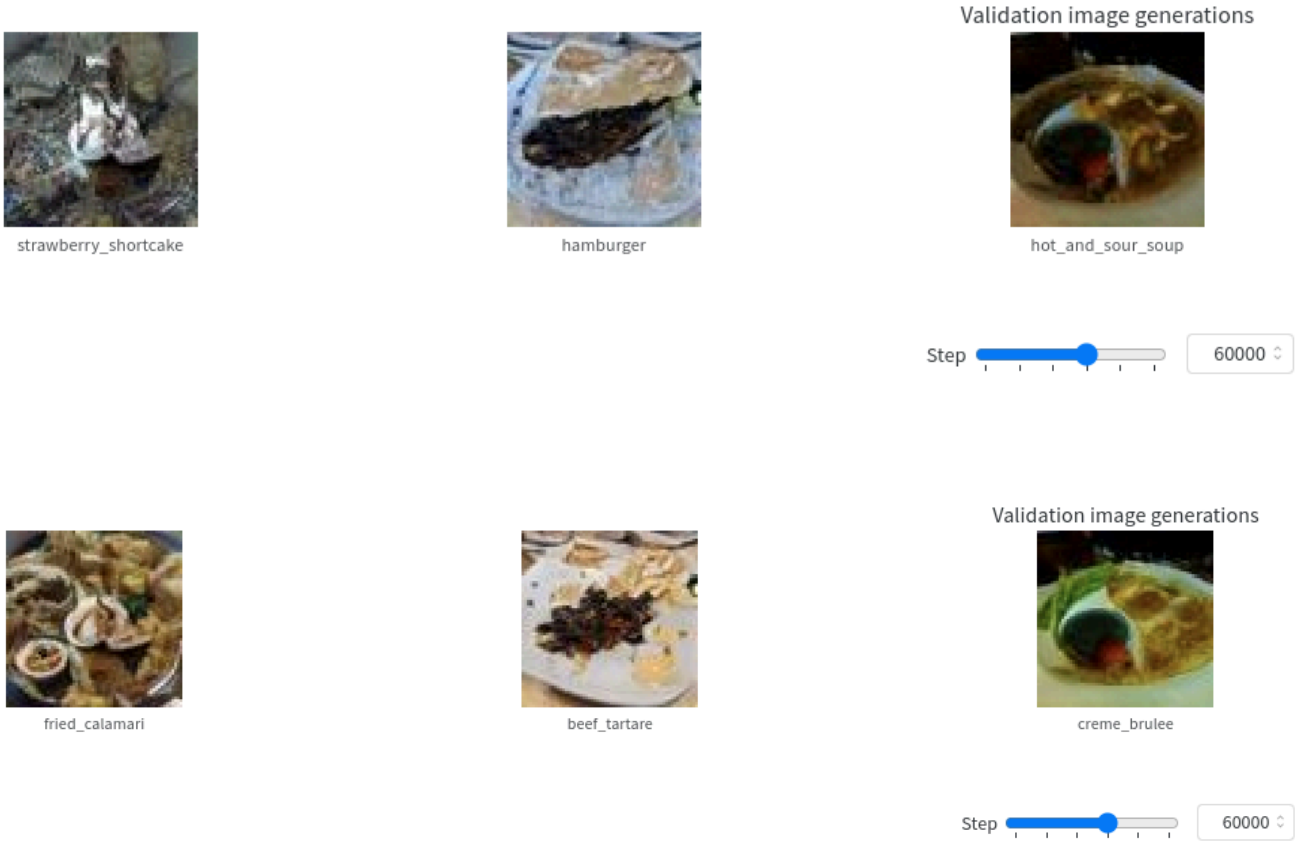

Рис 7. Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 60к итераций

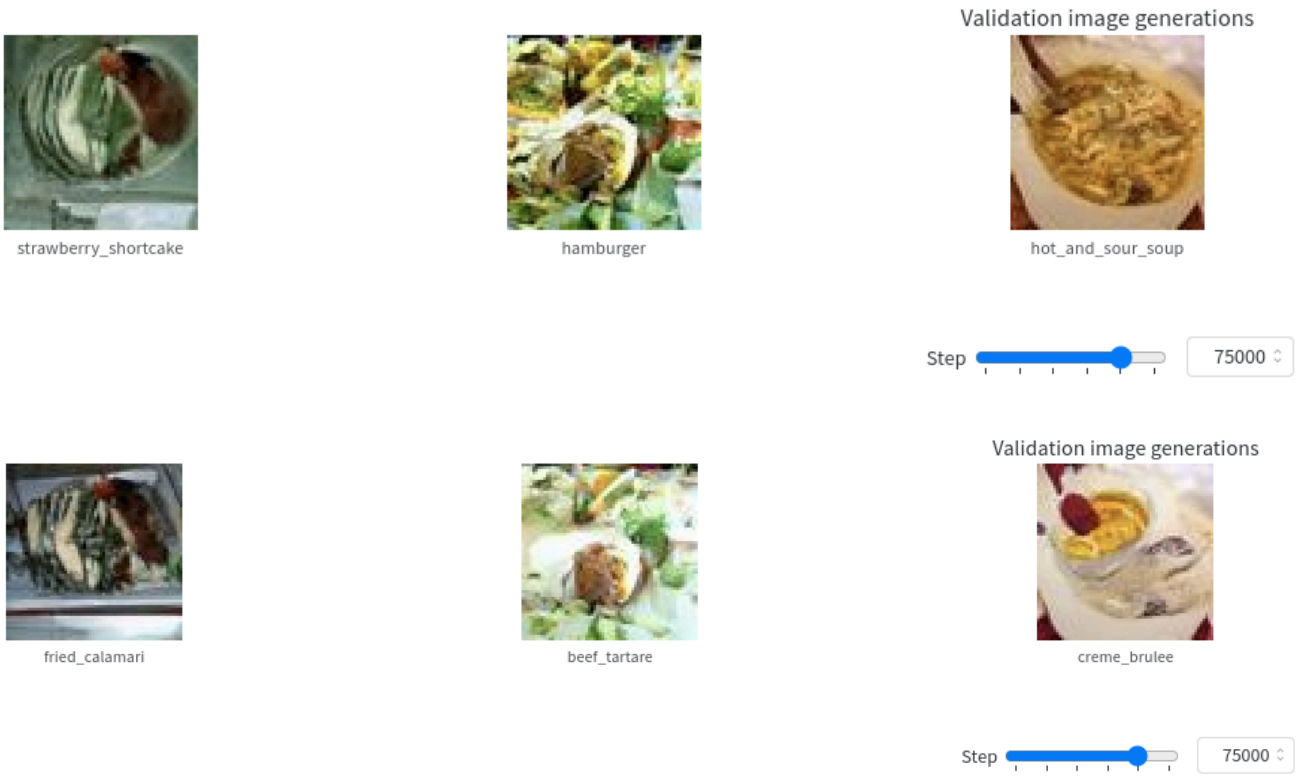

Рис 8. Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 75к итераций

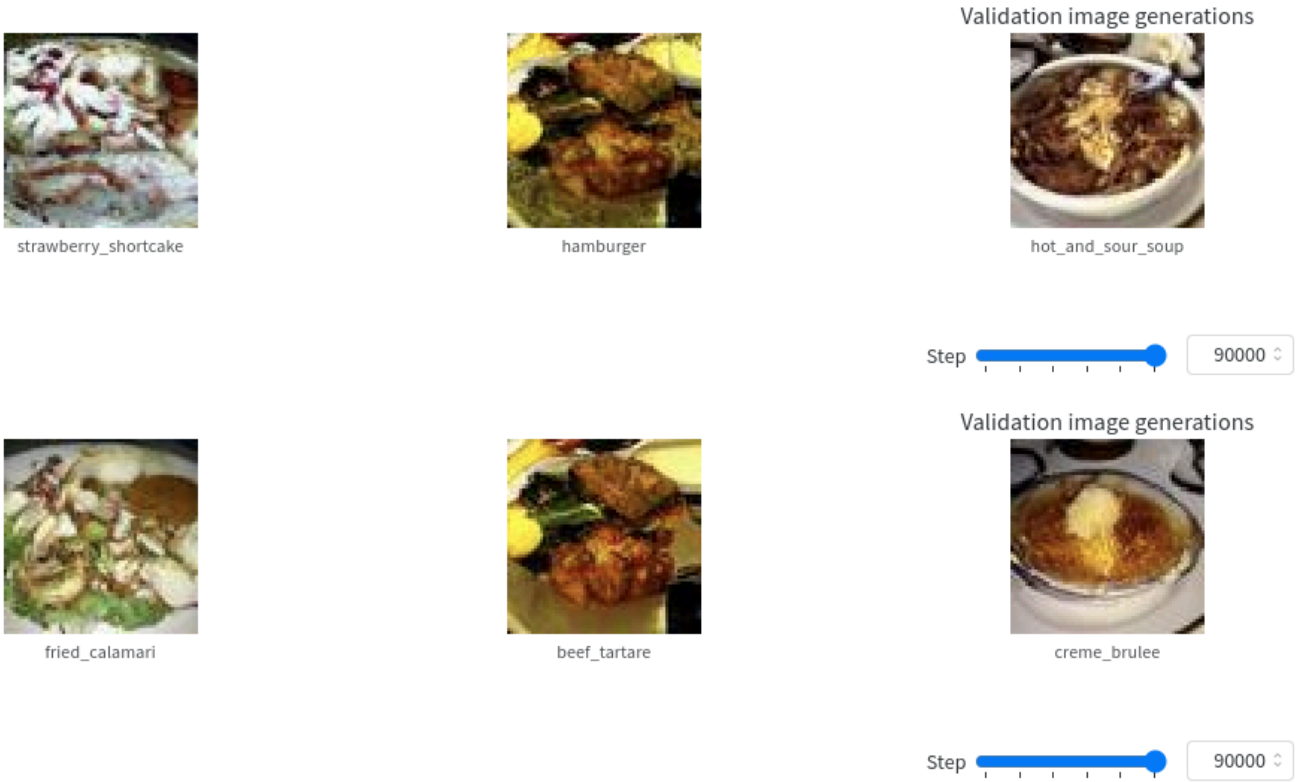

Рис 8. Изображения сгенерированные моделью MSE (верхний ряд) VLB + MSE (нижний ряд), обученной на 90к итераций

Детали обучения моделей

Обучение производилось в google collab на видеокарте L4. Каждая модель обучалась 90000 итераций. Размер батча 64 картинки.

В качестве оптимизатора был выбран Adam с дефолтными параметрами. FID считался каждые 15к итераций. LR - 0.004.

Ссылки и заимствования

1. Improved diffusion - https://arxiv.org/abs/2102.09672
2. DDPM - https://arxiv.org/abs/2006.11239
3. Stable diffusion - https://arxiv.org/abs/2112.10752
4. Diffusion Models Beat GANs on Image Synthesis - https://arxiv.org/abs/2105.05233
5. Код реализации “Improved diffusion” был позаимствован из https://github.com/openai/improved-diffusion