Generative Modeling by Estimating Gradients of the Data Distribution

Анищенко Илья, БПМИ 171

Вступление:

Два основных видна генеративных моделей:

- На основе вероятностных методов
- На основе состязательного обучения

Проблемы вероятностных моделей:
Использование специальных архитектур для построение нормализованной вероятностной модели
Использование суррогатных функций потерь

GANы же нестабильны из-за своей процедуры состязательного обучения.

Score-based generative modeling

Основные понятия

score: $\nabla_{\mathbf{x}} \log p(\mathbf{x})$

Score network: $\mathbf{s}_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^D$

Score matching for score estimation

$$\frac{1}{2}\mathbb{E}_{p_{\text{data}}}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\right\|_2^2\right]$$

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\operatorname{tr}(\nabla_{\mathbf{x}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})) + \frac{1}{2}\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})\right\|_2^2\right]$$

Однако score matching не масштабируется для глубоких сетей и высокоразмерных данных из-за вычисления следа.
Поэтому для этих задач используют модернизированные подходы

Denoising score matching

Распределение шума $q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})$

Зашумленные данные $q_\sigma(\tilde{\mathbf{x}}) \triangleq \int q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) p_{\text{data}}(\mathbf{x}) \mathrm{d}\mathbf{x}.$

Новый вид оптимизации $\frac{1}{2}\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})p_{\text{data}}(\mathbf{x})}[\|\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})\|_2^2].$

Получившийся score network $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}) = \nabla_{\mathbf{x}} \log q_\sigma(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$

Score network принимает такое значение, когда $q_\sigma(\mathbf{x}) \approx p_{\text{data}}(\mathbf{x}).$

Sliced score matching

Используем случайные проекции для аппроксимации следа, и как итог получаем:

$$\mathbb{E}_{p_{\mathbf{v}}} \mathbb{E}_{p_{\text{data}}} \left[\mathbf{v}^\intercal \nabla_{\mathbf{x}} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) \mathbf{v} + \frac{1}{2} \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) \right\|_2^2 \right]$$

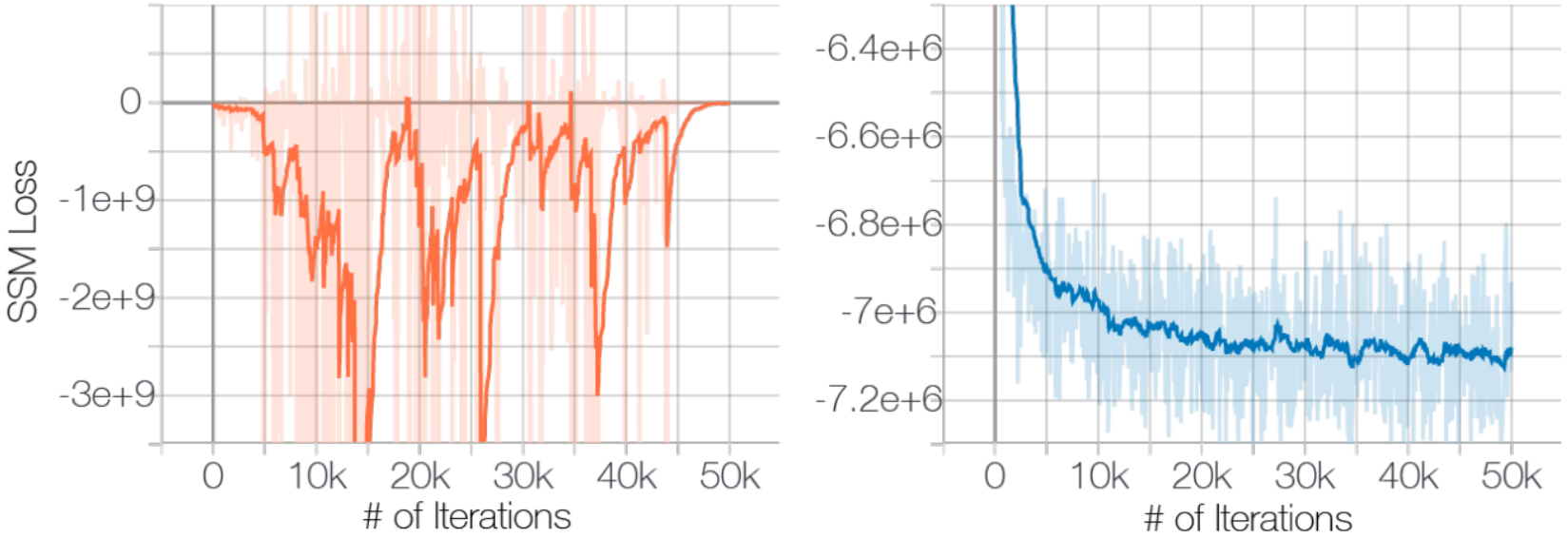
$\mathbf{v}^\intercal \nabla_{\mathbf{x}} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) \mathbf{v}$ Вычисляется методами авто-дифференцирования

Sampling with Langevin dynamics

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \frac{\epsilon}{2} \nabla_{\mathbf{x}} \log p(\tilde{\mathbf{x}}_{t-1}) + \sqrt{\epsilon}\, \mathbf{z}_t,$$

Проблемы данного score-based подхода

Обучение ResNet с SSM loss на CIFAR-10:



Слева: Sliced score matching (SSM) loss от номера итерации, без добавления шума в данные. Справа: тоже самое, но в данные добавлен шум - $\mathcal{N}(0, 0.0001)$.

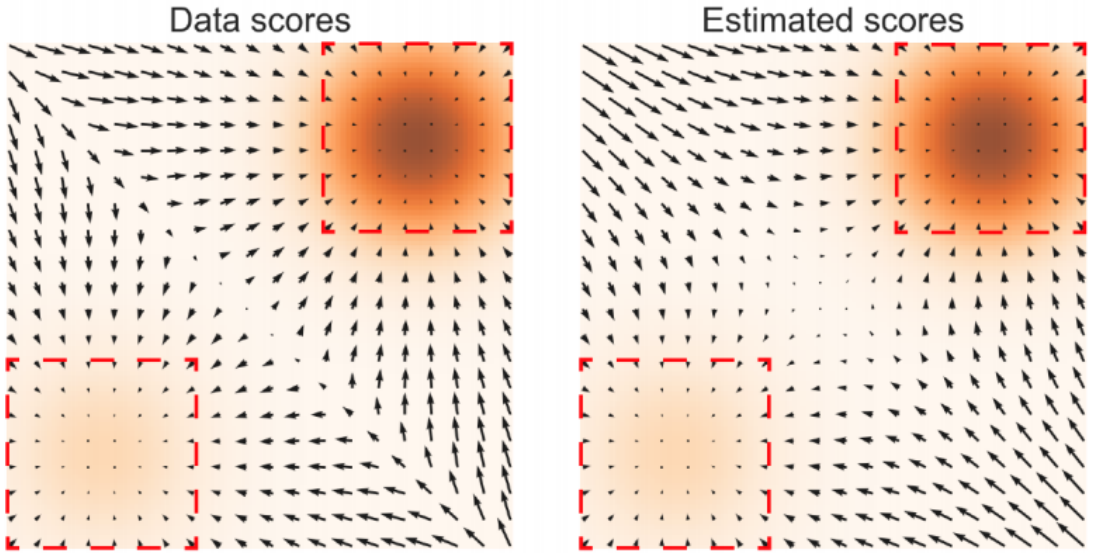
Гипотеза многообразия: данные в реальном мире имеют тенденцию концентрироваться на низкоразмерных многообразиях, встроенных в высокомерное пространство (иначе называемое окружающим пространством). Эта гипотеза эмпирически справедлива для многих наборов данных и стала основой многомерного обучения [3, 47]. В соответствии с гипотезой многообразия генеративные модели, основанные на баллах, столкнутся с двумя ключевыми трудностями.

Low data density regions

Inaccurate score estimation with score matching

$$p_{\text{data}} = \frac{1}{5}\mathcal{N}((-5,-5), I) + \frac{4}{5}\mathcal{N}((5,5), I)$$

$$\frac{1}{2}\mathbb{E}_{p_{\text{data}}}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|_2^2\right]$$

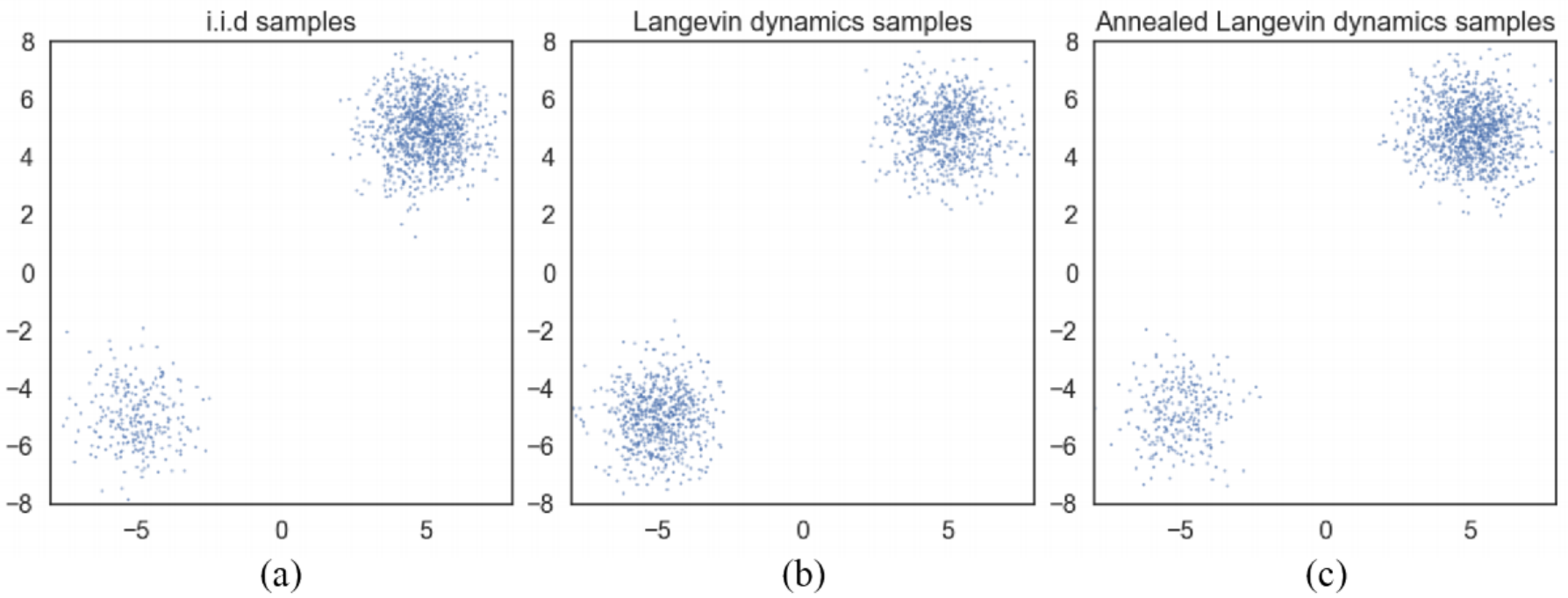


выборка $\{\mathbf{x}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} p_{\text{data}}(\mathbf{x})$

рассмотрим $\mathcal{R} \subset \mathbb{R}^D$ что $p_{\text{data}}(\mathcal{R}) \approx 0$ тогда $\{\mathbf{x}_i\}_{i=1}^N \cap \mathcal{R} = \varnothing$

Slow mixing of Langevin dynamics

$$p_{\text{data}}(\mathbf{x}) = \pi p_1(\mathbf{x}) + (1-\pi) p_2(\mathbf{x}) \qquad p_{\text{data}} = \frac{1}{5}\mathcal{N}((-5,-5), I) + \frac{4}{5}\mathcal{N}((5,5), I)$$



Сэмплы из семи распределений Гаусса с различными методами:

а) точная выборка

б) Сэмплы с использованием динамики Ланжевена

в) Сэмплы с использованием отоженной динамикой Ланжевена (про неё ниже)

Noise Conditional Score Networks: обучение и вывод

Можно заметить, что данные с добавленным случайным гауссовским шумом выдают нам распределение данных более поддающееся для score-based моделей.

- поскольку опорой нашего распределения гауссовского шума является все пространство, возмущенные данные не будут ограничены низкоразмерным многообразием, что устраняет трудности с гипотезой многообразия и делает оценку баллов хорошо определенной.

- большой Гауссов шум имеет эффект заполнения областей низкой плотности в исходном невозмущенном распределении данных, поэтому score matching может получить больше обучающего сигнала для улучшения оценки score.

Предлагаемые нововведения в текущую score-based модель:
Возмущение данных с использованием различных уровней шума
Одновременная оценка score, соответствующих всем уровням шума, путем обучения одной условной score network
Использованием отоженной динамики Ланжевена. Т.е. при её использовании для генерации сэмплов мы сначала используем оценки, соответствующие большому шуму, и постепенно переходим на результат генераций с меньшим шумом.

Noise Conditional Score Networks

$\{\sigma_i\}_{i=1}^{L} \quad \frac{\sigma_1}{\sigma_2} = \cdots = \frac{\sigma_{L-1}}{\sigma_L} > 1$

Геометрическая последовательность выбранных сигм

$q_\sigma(\mathbf{x}) \triangleq \int p_{\text{data}}(\mathbf{t}) \mathcal{N}(\mathbf{x} \mid \mathbf{t}, \sigma^2 I)$

Распределение зашумленных данных

Общий score network для всех уровней шума $\forall \sigma \in \{\sigma_i\}_{i=1}^{L} : \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, \sigma) \approx \nabla_{\mathbf{x}} \log q_\sigma(\mathbf{x})$

Learning NCSNs via score matching

$q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}} \mid \mathbf{x}, \sigma^2 I)$

$\nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) = -(\tilde{\mathbf{x}} - \mathbf{x})/\sigma^2$

$$\ell(\boldsymbol{\theta}; \sigma) \triangleq \frac{1}{2} \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \sigma) + \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2 \right]$$

$$\mathcal{L}(\boldsymbol{\theta}; \{\sigma_i\}_{i=1}^{L}) \triangleq \frac{1}{L} \sum_{i=1}^{L} \lambda(\sigma_i) \ell(\boldsymbol{\theta}; \sigma_i)$$

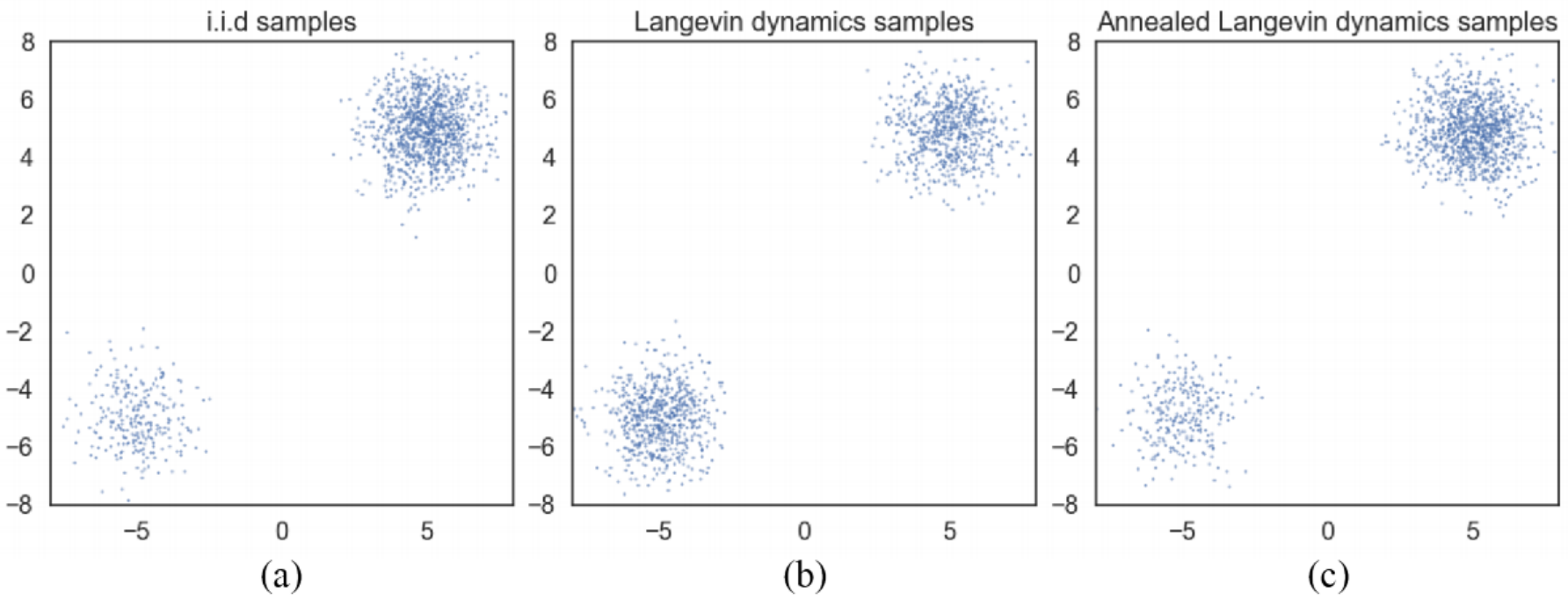
NCSN inference via annealed Langevin dynamics

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2 / \sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2} \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, \mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

NCSN inference via annealed Langevin dynamics



Сэмплы из семи распределений Гаусса с различными методами:

а) точная выборка

б) Сэмплы с использованием динамики Ланжевена

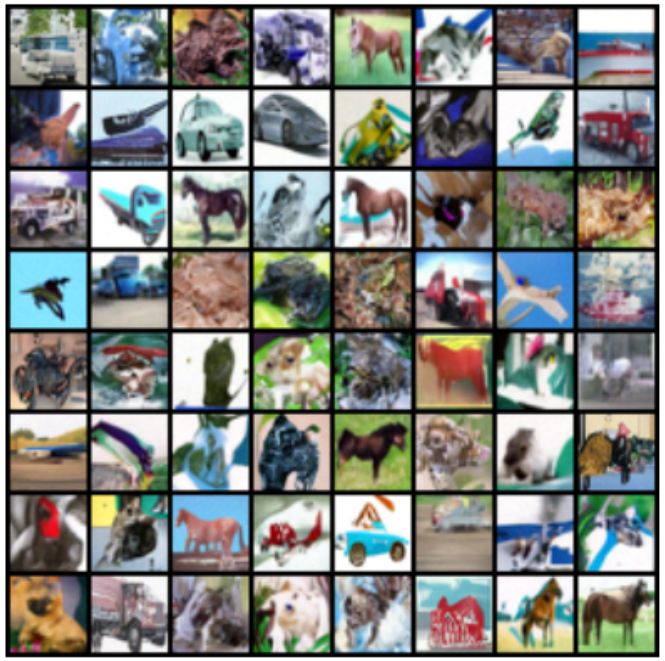
в) Сэмплы с использованием отоженной динамикой Ланжевена (про неё ниже)

Эксперименты:

Image generation

(a) MNIST

(b) CelebA

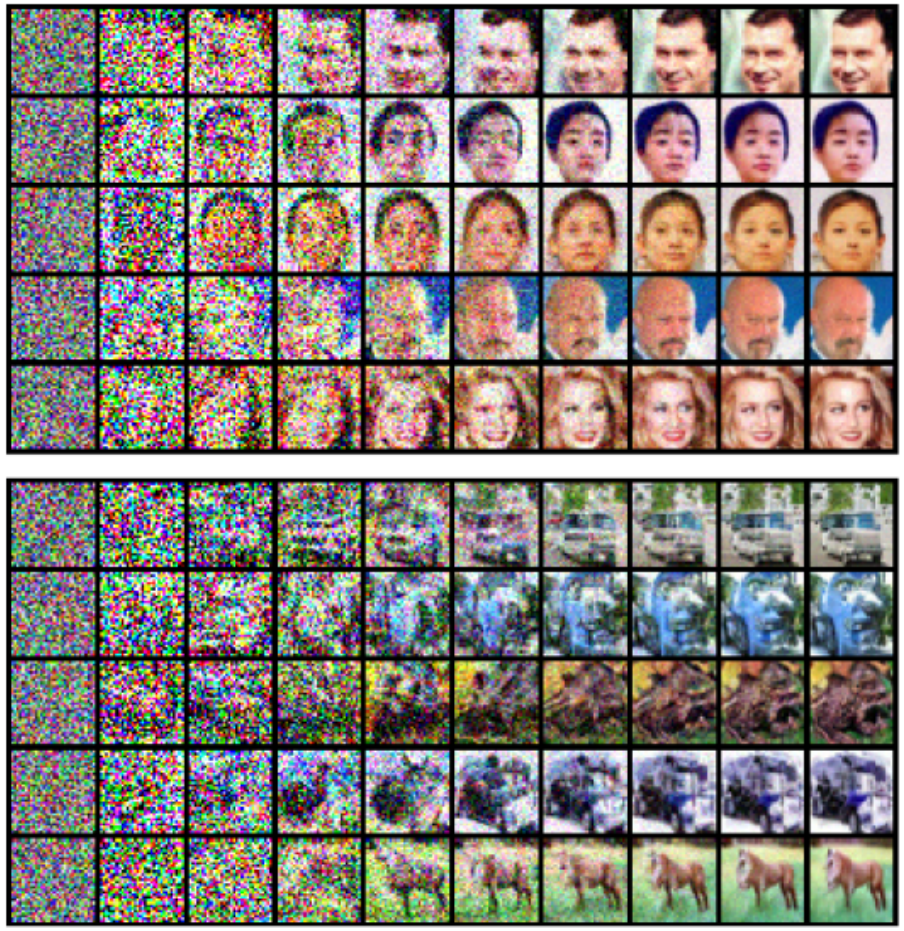


(c) CIFAR-10

Эксперименты:

Image generation

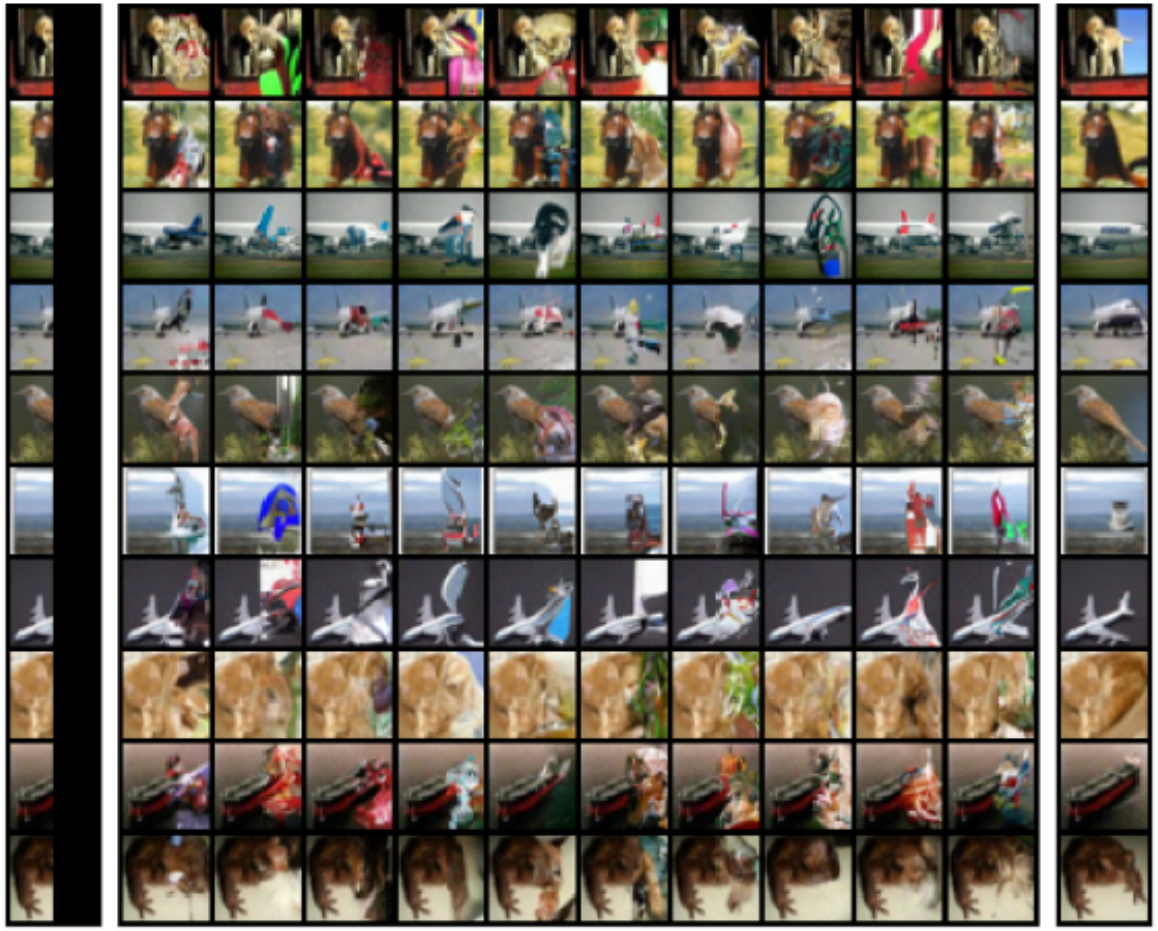
Model	Inception	FID
CIFAR-10 Unconditional		
PixelCNN [59]	4.60	65.93
PixelIQN [42]	5.29	49.46
EBM [12]	6.02	40.58
WGAN-GP [18]	$7.86 \pm .07$	36.4
MoLM [45]	$7.90 \pm .10$	**18.9**
SNGAN [36]	$8.22 \pm .05$	21.7
ProgressiveGAN [25]	$8.80 \pm .05$	-
NCSN (Ours)	$\mathbf{8.87} \pm .12$	25.32
CIFAR-10 Conditional		
EBM [12]	8.30	37.9
SNGAN [36]	$8.60 \pm .08$	25.5
BigGAN [6]	**9.22**	**14.73**



Переходы отоженной динамики Ланжевена между шумовыми распределениями данных. От большего шума к меньшему

Эксперименты:

Image inpainting



Заключение:

- Была предложена структура генеративного моделирования формата score-based, где сначала оцениваются градиенты плотности данных с score matching, а затем генерируются объекты с помощью динамики Ланжевена.

- Был проведен анализ нескольких проблем, с которыми сталкивается наивное применение этого подхода.

- Было предложено решать их путем обучения условно шумовых score сетей (NCSN) и генерировать выборки с отожженной динамикой Ланжевена.

- Такой подход не требует никакого состязательного обучения, использования MCMC во время обучения и никаких специальных архитектур моделей.

- Экспериментально было показано, что такой подход может генерировать высококачественные изображения, которые ранее были получены только с помощью лучших вероятностных моделей и GAN.

Вопросы:

- Какие нововведения предлагают авторы в своем подходе в задаче генерации?
- С какими основными проблемами сталкиваются авторы статьи в своей модели генерации?
- Как добавление гауссовского шума влияет на обучение модели?
- Кратко опишите алгоритм с динамикой отжига Ланжевена для генерации данных из обученной авторами модели NCSN