УДК 004.89
doi: 10.21685/2587-7704-2023-8-2-6

Open Access RESEARCH ARTICLE

# Персонализация генерации лиц с помощью модифицированной текстовой инверсии

***Григорий Борисович Лившиц***
Пензенский государственный университет, Россия, г. Пенза, ул. Красная, 40
kpebedkoz@gmail.com

**Аннотация.** Рассматривается модификация метода текстовой инверсии применительно к контексту персонализированной генерации лиц. Показано, что использование репараметризации объекта оптимизации (текстовых эмбеддингов) дает прирост в косинусной близости CLIP, в то время как использование улучшенной инициализации уменьшает косинусную близость CLIP, при этом сильно увеличивая схожесть сгенерированных лиц с целевой личностью.

**Ключевые слова**: текстовая инверсия, генерация лиц

**Для цитирования**: Лившиц Г. Б. Персонализация генерации лиц с помощью модифицированной текстовой инверсии // Инжиниринг и технологии. 2023. Т. 8 (2). С. 1–5. doi: 10.21685/2587-7704-2023-8-2-6

# Personalized generation of facial images using modified textual inversion

***Grigoriy B. Livshits***
Penza State University, 40 Krasnaya Street, Penza, Russia
kpebedkoz@gmail.com

**Abstract.** The present work is dedicated to the modification of textual inversion for personalized faces generation. It is shown that reparameterization of optimized objective (textual embeddings) boosts CLIP score, whereas modified initialization significantly increases ID score at the price of lower CLIP score.

**Keywords**: textual inversion, faces generation

**For citation**: Livshits G.B. Personalized generation of facial images using modified textual inversion. *Inzhiniring i tekhnologii = Engineering and Technology.* 2023;8(2):1–5. (In Russ.). doi: 10.21685/2587-7704-2023-8-2-6

## Введение

На сегодняшний день самым передовым методом генерации контента является использование диффузионных моделей, качество генераций которых превосходит генеративно-состязательные сети [1]. Особенно выдающиеся результаты демонстрируют диффузионные модели, способные генерировать изображения из текста: получающиеся результаты генерации разнообразны, имеют высокое качество и способны содержать в себе смесь разнообразных сложных концепций. Более того, последние работы по персонализации диффузионных моделей, такие как *Dreambooth* [2] и *Textual Inversion* [3], позволяют сгенерировать изображения по собственным концептам, например, сгенерировать собственную машину или памятную вещь в уникальных сценариях, формах и расцветках.

Наиболее популярным запросом для персонализации генераций является представление определенных людей в новой ситуации, например: человек хочет увидеть себя в образе супергероя или свой портрет в постмодернистском стиле. Приведенные выше методы обладают рядом недостатков в этом домене генерации лиц. Во-первых, результат текстовой инверсии очень часто не сохраняет мелкие особенности и детали концепта, что в контексте генерации конкретных лиц может приводить к identity *leak* – отличию сгенерированного лица от реального. Более того, оригинальный метод текстовой инверсии предполагает тысячи итераций оптимизации, что может занимать много времени и приводить к переобучению. Во-вторых, в результате файн-тюнинга диффузионной модели с помощью *Dreambooth*

она может быть переобучена, в результате чего модель хоть и сможет генерировать нужного человека, но не сможет инкорпорировать его в новые концепты; также каждая отдельная персонализация занимает несколько гигабайт памяти (вес оригинальной модели), что при большом количестве пользователей может оказаться очень затратным.

Целью данной работы является модификация оригинального метода текстовой инверсии для персонализации генераций лицевых изображений. Наш вклад заключается в следующем:

1) улучшенная инициализация параметров текстовой инверсии, специфическая для выбранного домена;

2) репараметризация обучаемых параметров текстовой инверсии, существенно снижающая эффект переобучения.

## Улучшенная инициализация текстовой инверсии

В классическом подходе текстовой инверсии эмбеддинги инициализируются либо наиболее редкими векторами из словаря текстовой модели (примеры соответствующих токенов – «*ljz*», «*sks*»), либо кратким описанием целевого объекта или домена (например, использование токена «*dog*» при персонализации генераций собак).

В данной работе были рассмотрены три способа инициализации текстовых векторов:

1) инициализация фразой «*human face*»;

2) инициализация двумя случайными векторами из векторизованных имен знаменитостей, которые присутствовали при обучении модели;

3) инициализация векторами, соответствующими самой похожей на целевую личность знаменитости.

В экспериментах эти походы именуются как «*face*», «*rand*2» и «*id*», соответственно.

Интуиция в подходе 2 и 3 заключается в следующем: уникальным идентификатором человека являются его имя и фамилия. Использование подхода 3 позволяет на первых же этапах оптимизации получать лица, похожие на целевую личность. Похожесть между личностями измерялась с помощью распознавания лиц.

## Репараметризация объекта оптимизации

В классической текстовой инверсии требуется найти такой вектор (текстовый эмбеддинг), при использовании которого диффузионная модель сможет генерировать изображения целевого домена, например, изображения определенного человека, объекта или стиля. Однако при наивной оптимизации текстовых эмбеддингов их норма сильно отличается от нормы векторов в словаре эмбеддингов текстовой модели, что может привести к неудовлетворительным генерациям целевого объекта: текстовое описание может либо не иметь должного эффекта, либо лицо будет сильно изменено и не будет соответствовать нужной личности. Такая же проблема наблюдается, например, при использовании инверсии изображений лиц в пространство *StyleGAN* [4].

Для борьбы с этой проблемой предлагается использовать в качестве параметров оптимизации ортонормированные векторы и логарифмы их нормы.

$$v = a^n \cdot \frac{d}{\|d\|}, \tag{1}$$

где $v$ – текстовый эмбеддинг;
$a$ – скалярный гиперпараметр – основание логарифма нормы;
$n$ – обучаемый логарифм нормы;
$d$ – обучаемый вектор.

Использование меньших оснований логарифма нормы позволяет получать эмбеддинги, максимально близкие по норме к изначальным, так как градиент функции потерь $L$ по $n$:

$$\nabla_n L = \nabla_v L \cdot \frac{d}{\|d\|} \cdot a^n \cdot In(a). \tag{2}$$

Так как нормы эмбеддингов текстовой модели изначально малы, оптимизация в логарифмическом пространстве позволяет сохранить норму обученных эмбеддингов, близкую к изначальным.

## Эксперимент

Для проведения экспериментов по модифицированной текстовой инверсии была взята модель *Stable Diffusion* 2.1-*base* – открытая диффузионная модель от компании StabilityAI. В качестве входных

данных было взято по 8 фотографий для 12 знаменитостей, при этом никакие изображения этих знаменитостей не использовались в обучении модели.

Для составления базы людей, которые были использованы в обучении диффузионной модели, нами были собраны изображения знаменитостей, которые затем были отфильтрованы по генерации этих знаменитостей моделью с помощью распознавания лиц.

Обучение проводилось в течение 400 итераций с помощью оптимизатора *Adam* [5] с $\beta 1 = 0{,}9$, $\beta 2 = 0{,}98$. При обучении без использования репараметризации добавлялась $l2$-регуляризация с коэффициентом 0,01. Обучение на каждом режиме производилось по 3 раза с использованием разных начальных состояний генератора случайных чисел.

Всего было проведено 9 экспериментов по инверсии: использовались все три инициализации, для каждой из которых проводилась одна инверсия без репараметризации и две с репараметризацией со значениями параметра $a = 2{,}71$ и $a = 2{,}0$.

В качестве функции потерь была выбрана взвешенная нижняя вариационная оценка [6]:

$$L_{simple} = \mathrm{E}_{t,x_0,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{a}_t}\,x_0 + \sqrt{1-\bar{a}_t}\,\varepsilon, v, t\right)\right\|^2\right], \quad (3)$$

где $x_0$ – экземпляры тренировочного набора данных;
$\varepsilon$ – экземпляр из стандартного нормального распределения;
$\varepsilon_\theta$ – шум, предсказанный диффузионной моделью;
$t$ – шаг диффузии;
$\alpha_t$ – коэффициент диффузии.

## Результаты

После обучения текстовых эмбеддингов по каждому из них было сгенерировано по 16 изображений для каждого из 25 заранее отобранных текстовых запросов. Генерация изображений производилась с помощью DDIM [7] с использованием 50 итераций.

В качестве метрики сохранения черт личности косинусная близость эмбеддингов модели распознавания лиц. Для контроля соответствия изображения текстовому описанию использовалась косинусная близость эмбеддингов модели CLIP [8]. Результаты экспериментов показаны на рис. 1.

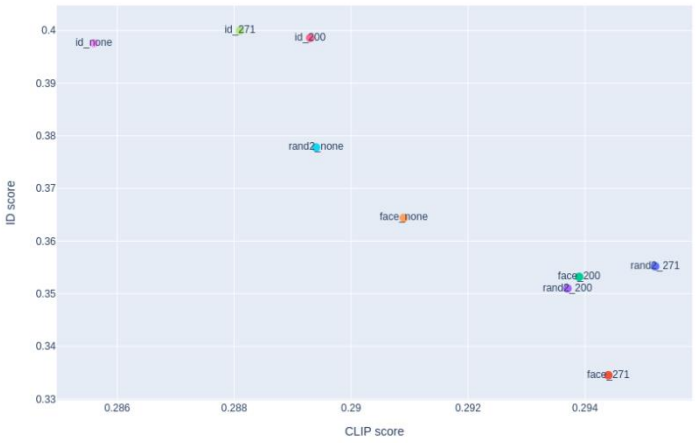


Рис. 1. Результаты экспериментов по инверсии.

Как видно из рис. 1, при всех инициализациях использования логарифма нормы дает прирост в близости CLIP, в то время как использование инициализации «*id*» дает лучшие результаты в плане сохранения индивидуальных черт лица. Это позволяет получать генерации, максимально похожие на изначальную личность, хотя оценка близости CLIP дает несколько худшие результаты. Для агрегированной оценки результатов предлагается следующая метрика:

$$score = \left(1 + score_{clip}\right) \cdot \left(1 + score_{id}\right), \quad (4)$$

где $score_{clip}$ – косинусная близость эмбеддингов CLIP;
$score_{id}$ – косинусная близость эмбеддингов модели для распознавания лиц.

Результаты расчета метрики показаны в табл. 1.

Таблица 1

Результаты расчета агрегированной метрики

| Логарифм нормы / Инициализация | «*face*» | «*rand*2» | «*id*» |
|---|---|---|---|
| none | 1,7611 | 1,7765 | 1,7978 |
| 2.71 | 1,7271 | 1,7546 | 1,8032 |
| 2 | 1,7500 | 1,7490 | 1,8029 |

Предположительно, использование других инициализаций дают худшие результаты из-за малого количества итераций, так как в пространстве эмбеддингов их стартовая точка находится дальше от оптимальной, чем при использовании инициализации «*id*». При использовании 400 итераций общее время инверсии составляет 3 минуты на *GeForce* RTX 3090 TI.

Примеры генераций показаны на рис. 2.

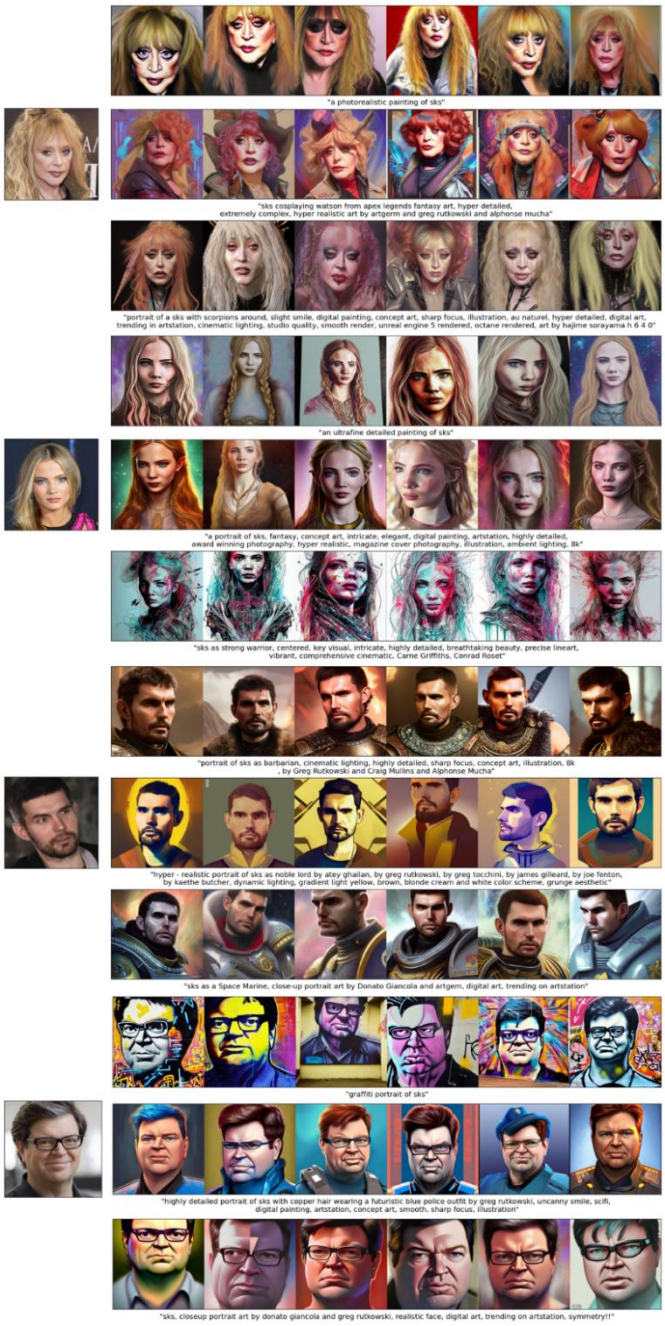


Рис. 2. Примеры генерации с использованием инициализации «*id*» и *a* = 2.0.
Вектор токена «*sks*» заменяется на обученный эмбеддинг

## Список литературы

1. Dhariwal P., Nichol A. Diffusion models beat GANs on image synthesis // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 8780–8794.
2. Ruiz N. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation // arXiv preprint arXiv:2208.12242. 2022.
3. Gal R. An image is worth one word: Personalizing text-to-image generation using textual inversion // arXiv preprint arXiv:2208.01618. 2022.
4. Tov O. Designing an encoder for StyleGAN image manipulation // ACM Transactions on Graphics (TOG). 2021. Vol. 40. № 4. P. 1–14.
5. Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014.
6. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 6840–6851.
7. Song J., Meng C., Ermon S. Denoising diffusion implicit models // arXiv preprint arXiv:2010.02502. 2020.
8. Radford A. et al. Learning transferable visual models from natural language supervision // International conference on machine learning. PMLR, 2021. P. 8748–8763.

## References

1. Dhariwal P., Nichol A. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*. 2021;34:8780–8794.
2. Ruiz N. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*. 2022.
3. Gal R. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*. 2022.
4. Tov O. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)*. 2021;40(4):1–14.
5. Kingma D.P., Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2014.
6. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*. 2020;33:6840–6851.
7. Song J., Meng C., Ermon S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*. 2020.
8. Radford A. et al. Learning transferable visual models from natural language supervision. *International conference on machine learning*. PMLR, 2021:8748–8763.

**Поступила в редакцию / Received** 15.05.2023

**Поступила после рецензирования и доработки / Revised** 17.06.2023

**Принята к публикации / Accepted** 30.07.2023