Image To Prompts

Kalimullin Timur

Abstract

Данный отчет содержит описание модели для решения задачи определения текста, использованного для генерации изображения. Код и данные представлены в GitHub репозитории - https://github.com/timllin/nlp-spring-term--2023.

1. Intorduction

Популярность задачи преобразования текстового описания в изображение существенно повлияло на развитие новой отрасли глубокого обучения. Это привело к активному исследованию и разработке новых алгоритмов и архитектур нейронных сетей, специализированных на генерации изображений из текстовых описаний. Тем не менее исследователи все еще пытаются определить, насколько текстовое описание и определенные тэги влияют на генерацию конечного изображения. Задача данного проекта - создать модель, которая позволит инвертировать процесс генерации изображения из текста.

1.1 Team

Автор проекта – Калимуллин Тимур, студент 1 курса магистратуры МФТИ.

2. Related work

Решение задачи генерации текста из изображения весьма трудоемкий процесс из-за своей мультимодальной природы.

1) Vision Pretraining. Обучение сложных моделей, таких как Transformers [1], на размеченных больших данных является популярной стратегией для многих задач компьютерного зрения. Также BEiT [2] предложил идею реализации BERT, но только для задач компьютерного зрения. Эти подходы не нашли решение для мультимодальной задачи Image to Text, однако были полезны в последующих исследованиях.

2) Vision-Language Pretraining. В последние годы произошло стремительное развитие моделей VLP(Vision – Language Pretraining), позволяющие кодировать и текстовые данные, и изображения. Более ранние работы, LXMERT [3], UNITER [4], использовали пред обученные модели

детекции для извлечения визуальных представлений изображений. Более поздние разработки, такие как ViLT [5] и VLMo [6], объединяют трансформеры (используемые и в задачах NLP и CV) для обучения мультимодального трансформера с нуля.

3) Новые исследования в данной области позволили успешно развить «image-text foundation models», которые включают в себе одновременно «vision» и «vision language pretraining» подходы. Так CLIP [7] представленный в 2021 году состоит из моделей кодирования изображений и текста, которые можно использовать для различных форм кросс-модального сравнения. LAION's OpenCLIP [8] – разработка группы независимых исследователей, которые улучшили модель CLIP, обучив ее на 2 миллиардах пар изображений/текстовых описаний. В 2022 году была представлена модель GIT, установившая SotA на многих бейнчмарках. Также в 2022 была представлена работа CoCa(Contrastive Captioners) [9], объясняющая как можно обучить модель генерировать текстовое описание из изображения автоматически. В 2023 году публикация работы [10] и датасета, позволяющий улучшить качество «image-text foundation» моделей.

3. Model Description

Данная модель должна предсказывать текстовое описание, которое было использовано для генерации изображения и конвертировать данное описание в эмбеддинг, представляющий собой вектор длиной равной 384.

Таким образом решено было использовать следующий подход, состоящий из следующих шагов:

- извлечь текст из изображения;
- закодировать текст в вектор эмбеддинга.

Для первого этапа будет использована модель CoCa OpenClip: coca_ViT-L-14, обученная на 1.8 миллиарда пар изображений/текстового описания. Данная модель представляет собой комбинацию двух популярных архитектур - Vision Transformer (ViT) и Language Transformer (L-14).

Для решения второго этапа задачи было решено использовать Sentence Transformer[9]. Была выбрана модель all-MiniLM-L12-v2, которая позволит конвертировать полученное текстовое описание изображения в вектор, длина которого равна 384. Архитектура all-MiniLM-L12 является легковесной и эффективной, позволяя модели генерировать высококачественные векторные представления предложений. Она состоит из 12 слоев трансформеров с эмбеддингами и вниманием. all-MiniLM-L12-v2 имеет примерно 66 миллионов обучаемых параметров.

4. Dataset

Данные в задачах глубокого обучения имеют ключевую роль влияющие на качество и точность обученной модели. Для данного проекта есть несколько путей создания тренировочного датасета: генерация собственных изображений с использованием «Stable Diffusion» или использование готового датасета. Было решено использовать готовый датасет. Данный датасет представляет собой пару: ссылку на изображение и текстовое описание(prompt, image). Датасет состоит из 100 строк.

	imgId	prompt
0	20057f34d	hyper realistic photo of very friendly and dys...
1	227ef0887	ramen carved out of fractal rose ebony, in the...
2	92e911621	ultrasaurus holding a black bean taco in the w...
3	a4e1c55a9	a thundering retro robot crane inks on parchme...
4	c98f79f71	portrait painting of a shimmering greek hero, ...

Таблица 1: Пример данных.

Для тестирования был использован представленный соревнованием датасет. Данные имеют идентичную структуру вышеописанного датасета и состоят из 7 пар изображений и текстовых описаний.

5. Experiment

5.1Metrics

Для анализа точности решения данной задачи будет рассчитываться средняя оценка косинусного сходства между предсказанными и фактическими векторами текстового описания, использованного для генерации изображения.

$$cosine\ similaty = S_c(A,B) \coloneqq \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

6. Results

В таблице 2 приведены итоговые результаты модели на имеющихся данных.

	Mean Cosine Similarity
Train Data	0.19
Test Data	0.17

Описанная модель генерирует из изображения текстовое описание. Пример, представлен ниже.

Рисунок 1: Сравнение текстового описания, использованного для генерации и текстового описания, полученного из модели.

Далее модель выдает уже вектор длинной 384, представляющий собой эмбеддинг данного описания.

7. Conclusion

Приведенный метод показал вышеописанные результаты в решении поставленной задачи. Так для дальнейшего развития модели и улучшения ее производительности, можно рассмотреть несколько следующих шагов:

1) Создание кастомного датасета: для расширения обучающих данных и адаптации модели к специфическим требованиям задачи можно создать собственный датасет. Это может включать генерацию собственных текстовых описаний и сопоставление их с соответствующими изображениями. Кастомный датасет позволит модели обучаться на более разнообразных и релевантных данных;

2) Использование аугментаций и трансформаций над изображениями: Аугментация изображений — это процесс применения различных преобразований к изображениям для создания новых вариаций данных. Это может включать изменение размера, поворот, смещение, изменение контраста и другие операции. Использование аугментаций позволит увеличить разнообразие данных и повысить качество модели.

3) Обучение более сложных моделей. Модель CoCa OpenClip: coca_ViT-L-14 является мощной, но дальнейшее развитие может включать использование более специфичных моделей. Это может быть достигнуто путем увеличения размера модели, добавления дополнительных слоев или экспериментов с различными архитектурами. Также возможно объединение

нескольких моделей в ансамбль, чтобы комбинировать их предсказания для получения лучшей общей производительности.

Вышеперечисленные шаги позволят улучшить модель, увеличить ее гибкость и повысить ее способность выделять более точные текстовые описания.

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
2. Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.
3. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.
4. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In ECCV, 2020.
5. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.
6. Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre- training with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358, 2021.
7. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.
8. M. Cherti et al., OpenCLIP, Reproducible scaling laws for contrastive language-image learning (2022)
9. J. Yu, CoCa: Contrastive Captioners are Image-Text Foundation Models(2022)
10. Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, Roy Schwartz: Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images(2023)