Робастый метод детекции машинносгенерированных изображений

A PREPRINT

Килинкаров Георгий
Кафедра анализа данных
Московский физико-технический институт(МФТИ)
Москва, Россия
kilinkarov.gv@phystech.edu

Даниил Дорин
Аффиляции
Адрес
email

Андрей Грабовой
Аффиляции
Адрес
email

1 Аннотация

В связи с улучшением качества машиносгенерированных изображений становится очень сложно отличать реальное изображение от сгенерированных. Существующие на данный момент решения имеют низкую обощающую способность. В этой статье рассматриваются разные модели, в том числе несвязанные с нейронными сетями. Также используется вся существующая информацию и модели, для подбора наилучшего решения. Дополнительно строится модель, которая сначала проверяет метод генерации, потом уже использует конкретную модель для этого метода генерации. Помимо этого, используются методы графических редакторов, на основе искуственного интеллекта.

Keywords Машинносгенерированные изображения

2 Введение

В современном мире в связи с развитием генераторов изображений человеческому глазу стало уже слишком сложно отличать настоящие изображение и машиносгенерированное. Ещё сложнее человеку отличить реальное изображение от реального, но с использованием графического редактора.[1] В связи с доступностью этих сервисов стали очень распространены разные виды мошенничества, использующие машиногенерацию. Таким образом задача детекции машинносгенерированных изображений стала очень важна.

На данный момент не существует общего подхода к решению этой задачи, устойчивого относисльно появления новых моделей. Например, появление диффузионных моделей генерации изображений свело сущесвтующие на тот момент методы к точности около 60 процентов[2]. Таким образом, существующие на данный момент методы имеют низкую обощающую способность. Актуальные научные статьи на эту тему можно поделить на три типа: построение устойчивой модели с помощью добавления новых типов генерации в фазу обучения[3, 1], решение задачи с помощью методов, не использующих AI (с помощью классических методов и рассмотрения спектра света)[4], создание новых более мощных датасетов для данный задачи[2, 5].

AI-модели обучается на всё более новых и новых датасетах, включая в себя новые способы генерации, создаются способы онлайн-обучения [1], что улучшает постепенно качество, но концептуально не отличается от предыдущих методов и не обеспечивает устойчивость в случае, если появится более

иновационный метод генерации. До появления диффузионных моделей высокое качество показывал метод, рассматривающий спектр по Фурье [4]. Но на диффузионных моделях не показывает уже высокого качества.

Таким образом, в этой статье проводится попытка объеденить существующие методы и найти новый способ детекции машиносгенерированных изображений. Новизна заключается в объединении методов и построении модели, предполагающей сначала тип генерации, а потом проверяющей на генерацию сгенерировано ли изображение уже непосредственно с предположением определенного типа генерации.

Преимущество этого подхода заключается в подборе оптимальной модели для конкретного класса генирации, проблема заключается в высокой цене ошибки: если произойдет ошибка в предсказании класса генерации, то будет использоваться заведомо плохо подходящая модель.

В качестве векторизатора мы используем предобученный CLIP [6], который используется во множестве разных исследований для разных целей и задач [7, 8, 9], в том числе используется в качестве векторизатора для задач классификации [8, 9]

3 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{\boldsymbol{x_i}, y_i\},\ i = 1, ..., N,$$

где $\boldsymbol{x_i} \in \mathbb{N}_0^{H \times W \times C}$ — изображение размера $H \times W \times C$, $y_i \in \{0, 1\}$.

Необходимо построить отображение $\boldsymbol{F} : \mathbb{N}_0^{H \times W \times C} \to \{0, 1\}$.

Для нахождения оптимального отображения $\boldsymbol{F}^*$ в классе моделей $\mathcal{F}$ используется Binary Cross-Entropy Loss (BCE):

$$\boldsymbol{F}^* = \arg \min_{\boldsymbol{F}^* \in \mathcal{F}} \text{BCE}(F).$$

4 Теория

Отображение $\boldsymbol{F} : \mathbb{N}_0^{H \times W \times C} \to \{0, 1\}$. представляет из себя композицию двух отображений: $\boldsymbol{F} = \boldsymbol{f} \circ \boldsymbol{g}$, где:

$$\boldsymbol{f} : \mathbb{N}_0^{H \times W \times C} \to \mathbb{R}^d \ - \text{векторизация изображения}$$

$$\boldsymbol{g} : \mathbb{R}^d \to \{0, 1\} \ - \text{классификатор}$$

В статье для обучения $\boldsymbol{F}$ обучается только голова классифкатора $\boldsymbol{g}$, а $\boldsymbol{f}$ фиксировано и не обучается. Для векторизатора $\boldsymbol{f}$ рассматривается CLIP [6] .

CLIP [6] - модель, обученная на изображении и его текстовом представлении. На выходе получаем размер 512. Эта модель является одним из state-of-art классификаторов, для многих задач достаточно сделать голову классификатора и получится высокая точность для классификатора.

Одна из основный идей обучения CLIP [6] состоит в том, что модель при обучении использует помимо изображения ещё и её текстовое представление, которое создаётся также моделью CLIP [6] в другой фазе рбчения. Кодировщик изображения и кодировщик текста обучаются совместно. На рисунке 1 представлен краткий процесс обучения. Для нашей задачи от CLIP [6] нам нужен кодировщик изрбражения. В течение обучения использовалась серия из 5 ResNet и 3 Vision Transformer. В качестве оптимизатора используется оптимизатор Adam [10] с регуляризатором decoupled weight decay [11], применяемой ко всем весам, которые являются усилителями или смещением, и затуханием скорости обучения по графику косинусов [12].

5 Вычислительный эксперимент

Целью вычислительного эксперемента является проверка качества двухступенчатой классификации.

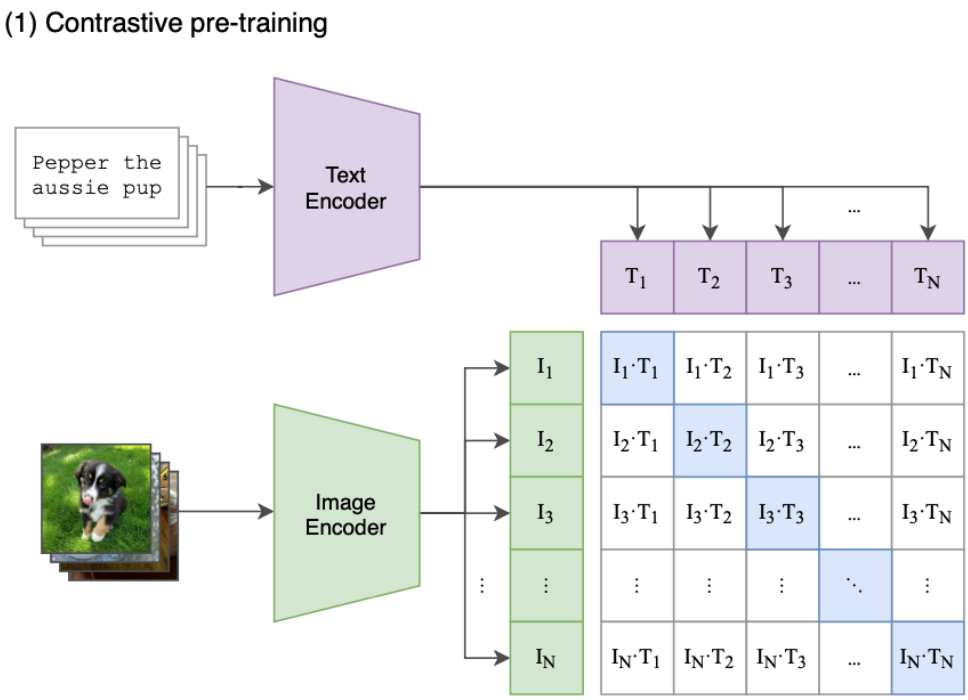

Figure 1: Описание подхода обучения CLIP [6]

В работе рассматривается датасет данных Artifact [13], для вычислительного эксперемента взято 13200 изображений. Датасет включается в себя реальные изображения и 25 методов генерации изображений, включая 13 GANs, 7 диффузионных, и 5 других методов генерации. Изображения по классам берутся так, чтобы их количество было одинаково для всех классов, кроме real(их очень много) и ddpm(их очень мало). Количество данных для классов на тестовой выборке можно посмотреть в таблице ??

Модель	real	stylegan1	stylegan2	stylegan3	big gan	pro gan	projected_gan	gau gan	star gan	gansformer	generative inpainting	mat	palette
Accuracy	919	53	141	54	53	51	53	53	53	53	53	54	53

Модель	tamping transformer	ddpm	lattent diffusion	stable diffusion	vq diffusion	glide	lama	denoising diffusion gan	face syntheyics	clips	cycle gan	sfhd	diffusion gan	Общее
Accuracy	54	6	53	53	53	53	54	53	53	53	53	53	53	2287

Table 1: Таблица количества данных для разных классов генерации на тестовой выборке. На тренировочном тоже самое, только данных больше.

В качестве основной модели будет использоваться CLIP [6]. Выборка была поделена на обучающую и тестовую в соотношении 70 на 30.

В работе рассматривается несколько разных моделей для сравнения и на основе этих эксперементов выбирается лучшая. За базовую модель была взята модель с CLIP [6] бинарным классификаором. Вторая модель к CLIP [6] добавляет два линейных слоя с функцией активацией Relu Один линейный слой с выходом 26(именно столько классов изображений существует в датасете). Третья модель является многоклассовым классификатором, но Test Accuracy считается для возможности сравнения как для бинарной модели. На Рисунке 2 представлена схема разных моделей.

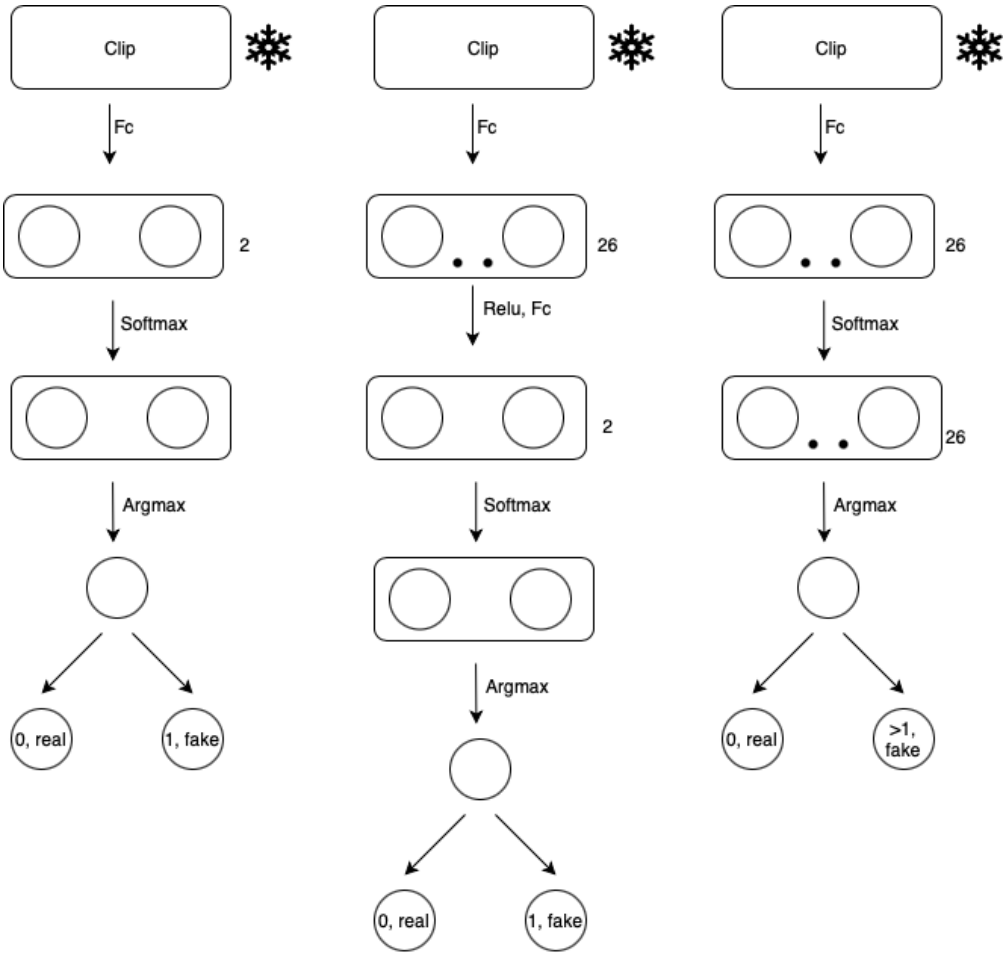

Figure 2: Общая схема для трех разных основных эксперементов: слева базовый, посередине модель с дополнительным слоем, а справа многоклассовая классификация. Значком снежинки обозначено то, что мы не обучаем CLIP [6], используем то, что уже существует.

Для оценки качества модели используются следующие оценочные метрики: Accuracy, Recall, Precision, ROC-AUC, PR-curve. Эти метрики помогают в проверке качества моделей и сравнения моделей между собой. Для многоклассовой классификацией также рассматривается confusion matrix для понимаю природы ошибок и слабых мест модели. Модели сравниваются по всем выше перечисленным метрикам, а также по графику обучения, включающие в себя 3 графика: график Loss от номера эпохи, Train, Test Accuracy от номера эпохи и время, затраченное на backward и forward от номера эпохи.

Обучение происходит с помощью оптимизатора Adam [10], с критерием CrossEntropy и 20 эпохами.

5.1 Базовая модель

Была обучена базовая модель по описанному выше общему принципу на 20 эпохах. Результаты Accuracy разных классов представлены в Таблице 2.

Для более полного понимания были построены Roc-Auc и PR-curve. Это представлено на Рисунке 3.

Модель	real	stylegan1	stylegan2	stylegan3	big gan	pro gan	projected_gan	gau gan	star gan	gansformer	generative inpainting	mat	palette
Accuracy	0.48	0.64	0.75	0.50	1.00	0.78	0.96	1.00	0.94	0.92	0.75	0.74	0.66

Модель	tamping transformer	ddpm	lattent diffusion	stable diffusion	vq diffusion	glide	lama	denoising diffusion gan	face syntheyics	clips	cycle gan	sfhd	diffusion gan	Общее
Accuracy	0.703	1.00	0.91	0.62	0.92	0.88	0.87	0.54	1.00	1.00	0.98	0.98	0.84	0.69

Table 2: Таблица accuracy для разных классов. Верным классом для класса генерации здесь мы считаем, если модель предсказала, что это изображение является машинносгенерированным.

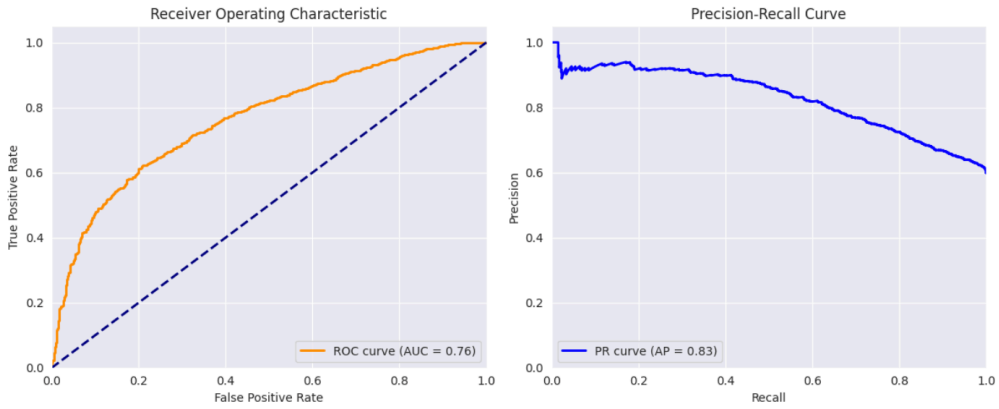

Figure 3: Roc-Auc и PR-curve для базовой модели

Помимо всего прочего был проведен анализ изображений, на которых наша базовая модель ошибается для учета этих ошибок и доработок. Рисунок 4 демонстрирует эти ошибки.

5.2 Модель с дополнительным слоем

Также было проведено сравнение первых двух моделей. Сравниваем по 3 графикам: график Loss от номера эпохи, график Accuracy от номера эпохи и время на проходы от номера эпохи.
Рисунок 5 показывает насколько добавление слоя улучшело модель при таком же времени работы.

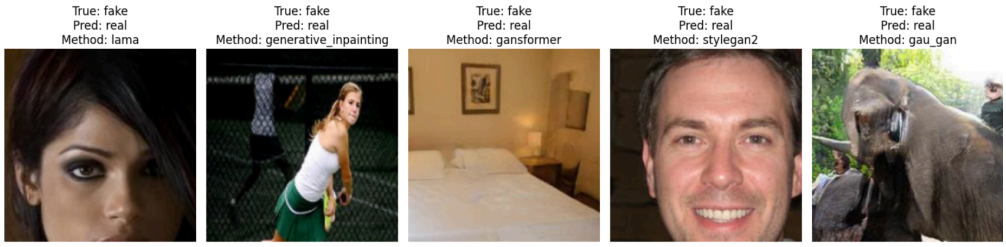

Figure 4: Примеры ошибок базовой модели

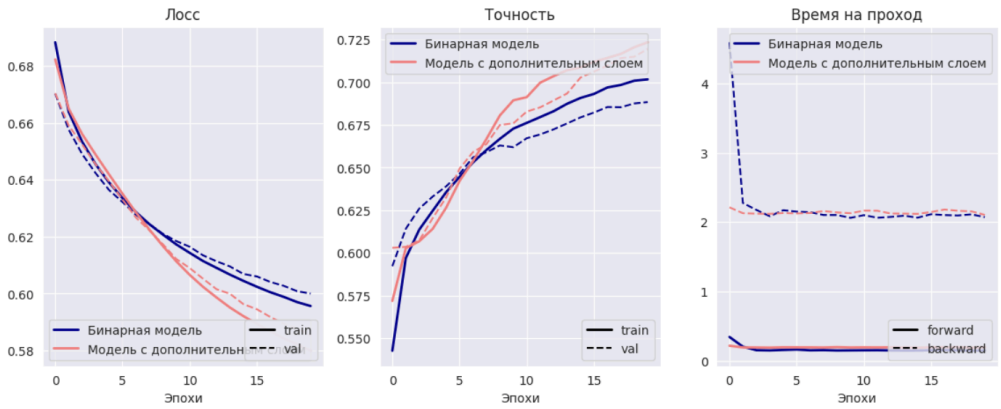

Figure 5: Сравнение первых двух моделей

5.3 Модель многоклассовой классификации

Модель многоклассовой классификации обучается на 5 эпохах с критерием - CrossEntropy с весами. В анализе считается, что модель предсказала правильно в случае, если реальное предсказано моделью как реальное, а сгенерированное как сгенерированное без учета конкретных классов генерации. Таки образом мы получаем сопостовимые для анализа модели. На Рисунке 6 представлено сравнение моделей. Как видим результат неудовлитворительный, так как модель многоклассовой классификации практически не обучилась. Для анализа проблем была рассмотрена confusion matrix (Рисунок 7).

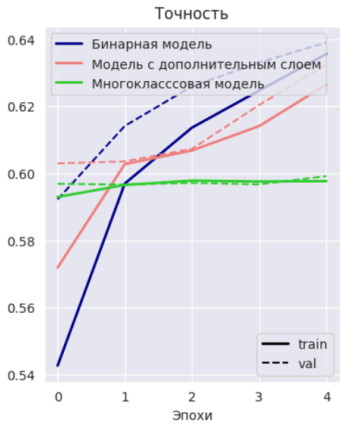

Figure 6: Сравнение первых трех моделей

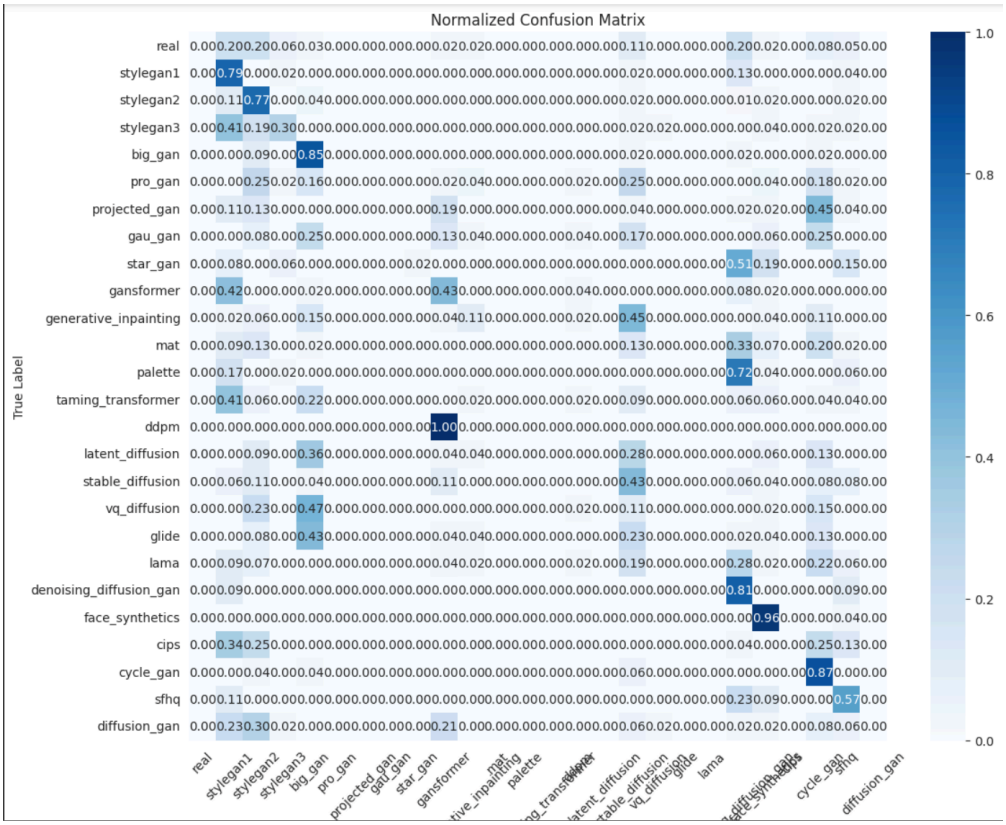

Figure 7: Confusion matrix для модели многоклассовой классификации.

References

[1] Oliver Wang Richard Zhang David C. Epstein, Ishan Jain. Online detection of ai-generated imagesonline detection of ai-generated images. 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/DFAD/html/Epstein_Online_Detection_of_AI-Generated_Images__ICCVW_2023_paper.html.

[2] Qiangyu Yan Xudong Huang Guanyu Lin Wei Li-Zhijun Tu Hailin Hu Jie Hu Yunhe Wang Mingjian Zhu, Hanting Chen. Genimage: A million-scale benchmark for detecting ai-generated image. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf.

[3] Tam V. Nguyen Samah S. Baraheem. Ai vs. ai: Can ai detect ai-generated images? 2023. URL https://www.mdpi.com/2313-433X/9/10/199.

[4] Matthias Nießner Luisa Verdoliva Davide Cozzolino, Giovanni Poggi. Zero-shot detection of ai-generated images. 2024. URL https://arxiv.org/abs/2409.15875.

[5] Ahmad Lotfi Jordan J. Bird. Image classification and explainable identification of ai-generated synthetic images. 2024. URL https://ieeexplore.ieee.org/abstract/document/10409290.

[6] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision. 2021. URL https://arxiv.org/pdf/2103.00020.

[7] Kerem Turgutlu Marcos V. Conde. Clip-art: Contrastive pre-training for fine-grained art classification. 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/CVFAD/papers/Conde_CLIP-Art_Contrastive_Pre-Training_for_Fine-Grained_Art_Classification_CVPRW_2021_paper.pdf.

[8] Xiaoguang Li Xiaofeng Wang Rabab Abdelfattah, Qing Guo and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Abdelfattah_CDUL_CLIP-Driven_Unsupervised_Learning_for_Multi-Label_Image_Classification_ICCV_2023_paper.pdf.

[9] Hao Tan Mohit Bansal Anna Rohrbach† Kai-Wei Chang‡ Zhewei Yao†and Kurt Keutzer† Sheng Shen, Liunian Harold Li. How much can clip benefit vision-and-language tasks? 2021. URL https://arxiv.org/pdf/2107.06383.

[10] Jimmy Lei Ba Diederik P. Kingma. Adam: A method for stochastic optimization. 2015. URL https://arxiv.org/pdf/1412.6980.

[11] Ilya Loshchilov Frank Hutter. Decoupled weight decay regularization. 2019. URL https://arxiv.org/pdf/1711.05101.

[12] Ilya Loshchilov Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2017. URL https://arxiv.org/pdf/1608.03983.

[13] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2200–2204, 2023. doi:10.1109/ICIP49359.2023.10222083.