**УДК 004.8**

# Создание алгоритма для обнаружения сгенерированных фрагментов на изображениях

**Масалимова А.А.** (ИТМО)

**Научный руководитель – старший научный сотрудник Ходненко И.В.** (ИТМО)

**Введение.** С развитием технологий искусственного интеллекта стало возможным создание фальшивых изображений, которые выглядят настолько реалистично, что зачастую их невозможно отличить от настоящих. Сгенерированные изображения (deepfake) активно применяются в различных областях, таких как цифровое искусство, маркетинг, киноиндустрия и даже на платформах социального взаимодействия. Это создает ряд серьезных угроз и этических проблем, связанных с подделкой изображений. В связи с этим разработка технологий автоматического обнаружения сгенерированных изображений становится важной задачей для обеспечения безопасности.

**Основная часть.** Генерация изображений — это процесс создания искусственных изображений с помощью различных методов ввода, например текста, эскиза, аудио или другого изображения [1]. Этот процесс используется во многих приложениях, таких как генерация произведений искусств [2], редактирование фотографий [3,4], наработка фотографий [5,6] и автоматизированный дизайн [7]. DeepFakes появились всего пару лет назад, но за это время эта технология успела усовершенствоваться и представляет большую угрозу. Подделки DeepFake могут быть созданы на основе большого количества изображений, доступных в интернете, поэтому создаются различные методы глубокого обучения для обнаружения поддельных изображений, таких как рекуррентная нейронная сеть (RNN), сверточная нейронная сеть (CNN) или длинная краткосрочная память (LSTM). А современные подходы к созданию Deepfake преимущественно базируются на архитектуре генеративно-состязательных сетей (GAN) [8].

Методы распознавания изображений можно разделить на два типа: один из них заключается в выявлении дефектов визуальных артефактов (например, в глазах, зубах и контурах лиц), а другой — это разработка модели глубокой нейронной сети для достижения дискриминации сгенерированных лиц. Статьи [9, 10, 11, 12] относятся к первой категории, упомянутой выше, в которых используется информация самого сгенерированного изображения для извлечения признаков. Работы [13, 14, 15, 16] принадлежат к категории проектирования моделей нейронных сетей, которые используются для реализации классификации сгенерированных лиц.

В рамках научно-исследовательской работы первоначально был проведен анализ артефактов изображений, со сгенерированными лицами людей. Так, вывод аналитик по каналам, подсчет статистик для цветовых схем, анализ насыщенности «чистых» цветов, вывод градиентов и построение распределения Фурье, не позволило однозначно классифицировать изображения, поэтому для создания качественного детектора была разработана модель нейронной сети.

В ходе разработки модели использовались данные, которые были созданы с помощью StyleGAN и StyleGAN2. Общий датасет состоит из 140 000 фальшивых и 70 000 реальных изображений. Для анализа всех возможных лиц на изображении была использована модель YOLA, которая позволила распознать и вырезать лицо человека. Для классификации изображений, представляющих реальные и сгенерированные лица, были протестированы различные нейронные сети. Каждая модель имеет свою уникальную архитектуру, которая адаптируется к задаче бинарной классификации, так были использованы модели: VGG19, ResNet, DenseNet, MobileNet, EfficientNet-V2-s.

Результаты обучения моделей продемонстрировали эффективность современных архитектур по сравнению с устаревающей VGG19. EfficientNet, DenseNet201 и ResNet101 показывают лучшие результаты по точности, при этом имеют умеренные вычислительные затраты. MobileNet является предпочтительной для задач с ограниченными ресурсами, так как

обеспечивает приемлемую точность при минимальных затратах. Также используемая техника визуализации Grad-CAM позволила рассмотреть тепловые карты, которые подчёркивают области изображения, наиболее значимые для предсказаний модели.

**Выводы.** Разработана модель нейронной сети, которая по результата теста на данных не из обучающего набора, позволяет однозначно определить является ли лицо на изображении сгенерированным или настоящим.

**Список использованных источников**:


1. Baraheem, S.S.; Le, T.-N.; Nguyen, T.V. Image synthesis: A review of methods, datasets, evaluation metrics, and future outlook. Artif. Intell. Rev. 2023, 56, 10813–10865.
2. Elgammal, A.; Liu, B.; Elhoseiny, M.; Mazzone, M. CAN: Generating 'art' by learning about styles and deviating from style norms.
3. Chen, J.; Shen, Y.; Gao, J.; Liu, J.; Liu, X. Language-Based Image Editing with recurrent attentive models.
4. Yan, Z.; Zhang, H.; Wang, B.; Paris, S.; Yu, Y. Automatic photo adjustment using deep neu-ral networks.
5. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with con-textual attention.
6. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. EdgeConnect: Generative image inpainting with adversarial edge learning.
7. Thaung, L. Advanced Data Augmentation: With Generative Adversarial Networks and Computer-Aided Design. 2020, Dissertation.
8. Guarnera L, Giudice O, Guarnera F, Ortis A, Puglisi G, Paratore A, Bui LMQ, Fontani M, Coccomini DA, Caldelli R, Falchi F, Gennaro C, Messina N, Amato G, Perelli G, Concas S, Cuccu C, Orrù G, Marcialis GL, Battiato S. The Face Deepfake Detection Challenge. J Imaging. 2022 Sep 28;8(10):263.
9. McCloskey, S.; Albright, M. Detecting Gan-generated imagery using saturation cues. In Proceedings of the IEEE International Conference on Image Processing (ICIP) 2019, Bordeaux, France, 16–19 October 2019; pp. 4584–4588.
10. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the Winter Applications of Computer Vision Workshops (WACVW) 2019, Seoul, Korea, 27 October–2 November 2019; pp. 83–92.
11. Zhang, X.; Karaman, S.; Chang, S.F. Detecting and simulating artifacts in GAN fake images. In Proceedings of the IEEE International Workshop on Information Forensics and Security 2019, Delft, The Netherlands, 9–12 December 2019; pp. 1–7.
12. Carvalho, T.; De Rezende, E.R.S.; Alves, M.T.P. Exposing computer generated images by eye's region classification via transfer learning of VGG19. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA) 2017, Cancun, Mexico, 18–21 December 2017; pp. 866–870.
13. Mo, H.; Chen, B.; Luo, W. Fake faces identification via convolutional neural network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security 2018, Innsbruck, Austria, 20–22 June 2018; pp. 43–47.
14. Minh Dang, L.; Ibrahim Hassan, S.; Im, S.; Lee, J.; Lee, S.; Moon, H. Deep Learning Based Computer Generated Face Identification Using Convolutional Neural Network. Appl. Sci. 2018, 8.
15. Nataraj, L.; Mohammed, T.M.; Manjunath, B.S.; Chandrasekaran, S.; Flenner, A.; Bappy, J.H. Detecting GAN generated Fake Images using Co-occurrence Matrice. Electron. Imaging 2019, 2019, 532-1–532-7.
16. Zhuang, Y.X.; Hsu, C.C. Detecting generated image based on a coupled network with two-step pairwise learning. In Proceedings of the IEEE International Conference on Image Processing (ICIP) 2019, Taipei, Taiwan, 22–29 September 2019; pp. 3212–3216.