УДК 004.056.57 : 004.852

ГИБРИДНАЯ АРХИТЕКТУРА CNN-KAN ДЛЯ ОБНАРУЖЕНИЯ СГЕНЕРИРОВАННЫХ ИЗОБРАЖЕНИЙ

Кудинов Владимир Вячеславович, Салита Даниил Сергеевич

Алтайский государственный университет, Барнаул
gifted2002@mail.ru; d.s.salita@gmail.com

HYBRID CNN-KAN ARCHITECTURE FOR DETECTING GENERATED IMAGES

Kudinov Vladimir V., Salita Daniil S.

Altai State University, Barnaul
gifted2002@mail.ru; d.s.salita@gmail.com

Аннотация. В статье рассматривается разработка гибридной архитектуры нейронной сети CNN-KAN, предназначенной для обнаружения сгенерированных изображений. Проблема детекции синтетического визуального контента становится все более актуальной на фоне бурного роста технологий генеративного ИИ. В работе описаны принципы функционирования сверточных нейронных сетей (CNN) и сетей Колмогорова–Арнольда (KAN), приведено обоснование их объединения в единую гибридную модель. Экспериментально показано, что CNN-KAN показывает более высокий результат чем традиционные CNN-MLP архитектуры по метрикам Accuracy, Precision, Recall, F1-score и ROC-AUC. Предложенная архитектура сочетает способность CNN выделять локальные признаки с возможностью KAN обобщать глобальные закономерности, что обеспечивает высокую точность классификации. Результаты исследования подтверждают эффективность гибридного подхода и перспективность его применения в задачах детектирования поддельных изображений.

Ключевые слова: дипфейк, сгенерированные изображения, машинное обучение, глубокое обучение, гибридная архитектура

Abstract. The article discusses the development of a hybrid CNN-KAN neural network architecture designed for detecting generated images. The problem of detecting synthetic visual content is becoming increasingly important due to the rapid growth of generative AI technologies. The paper describes the principles of convolutional neural networks (CNNs) and Kolmogorov-Arnold networks (KANs), and provides a rationale for combining them into a single hybrid model. It has been experimentally shown that CNN-KAN performs better than traditional CNN-MLP architectures in terms of Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. The proposed architecture combines the ability of CNN to extract local features with the ability of KAN to generalize global patterns, resulting in high classification accuracy. The research findings demonstrate the effectiveness of the hybrid approach and its potential for use in fake image detection tasks.

Keywords: deepfake, generated images, machine learning, deep learning, hybrid architecture

Для цитирования: *Кудинов В.В., Салита Д.С.* Гибридная архитектура cnn-kan для обнаружения сгенерированных изображений// Проблемы правовой и технической защиты информации. 2025. № 13. С. 17–22.

For citation: *Kudinov V.V., Salita D.S.* Hybrid cnn-kan architecture for detecting generated images. *Legal and Technical Problems of Information Security*. 2025. No. 13. P. 17–22.

Развитие алгоритмов генеративного машинного обучения за последнее десятилетие привело к появлению инструментов, способных создавать фотореалистичные изображения и видео, которые порой трудно отличить от подлинных. С одной стороны, это открыло новые горизонты для искусства, киноиндустрии и научных симуляций; с другой — породило серьезные угрозы, связанные с возможностью подделки визуального контента, дискредитации личности и распространения дезинформации. Определение генеративной модели звучит так: генеративные модели — это статистические модели, предназначенные для генерации данных. Их задача состоит в том, чтобы преобразовать шум в репрезентативную выборку данных.

Простые способы обнаружения такого рода изображений не справляются против все более качественных фейков, создаваемых современными генеративными сетями [13]. Поэтому на первый план выходят интеллектуальные методы на основе нейронных сетей. Концепция такого подхода состоит в обучении нейросетей классификаторов для обнаружения сгенерированных изображений.

CNN — сверточная нейронная сеть. CNN уже доказали свою эффективность в задачах распознавания лиц и объектов, а их способность анализировать изображение на разных иерархических уровнях делает их ключевым инструментом в выявлении визуальных фальсификаций. Достаточно обученная сеть такого типа с большой точностью выявляет сгенерированные изображения, но все же ошибается, в частности по причине ограничения архитектуры.

В связи с этим возникло предложение — комбинировать проверенную и надежную CNN с недавно предложенной архитектуры Kolmogorov–Arnold Networks (KAN), что позволяет объединить преимущества CNN с новым подходом к аппроксимации сложных зависимостей. Благодаря радикальному переосмыслению архитектуры, KAN обладают большей выразительной способностью при меньшем числе параметров, в контексте распознавания поддельных изображений это означает, что CNN-KAN модель способна уловить тонкие, нелинейные отличия фейка от подлинного изображения, потенциально избегая переобучения на специфичные артефакты генератора, так же увеличить точность распознавания.

Один из подходов заключается в замене традиционных полносвязных слоев CNN на слои KAN. В такой гибридной модели сверточные слои сначала выделяют локальные особенности изображения, а затем классификатор на основе KAN принимает эти признаки на вход и выдает решение к какому классу относится изображение, реальному или сгенерированному. В задачах детекции фейковых изображений: KAN-классификатор на выходе CNN улучшает способность модели обобщать, благодаря тому, что обучаемые активации гибче подстраиваются под распределение данных, чем фиксированные функции вроде ReLU. Более того, KAN-входы можно регуляризировать, что снижает риск переобучения на ограниченном наборе известных фейков и повышает устойчивость к новым типам фальсификаций. Другими словами, CNN-KAN архитектура способна уловить более общие закономерности, присущие поддельным изображениям и сохранить точность классификации, даже когда злоумышленники применяют новые методы генерации.

По мере распространения и развития ИИ-генерируемого контента эти инструменты продолжат становится все более эффективными на фоне простейших и устаревающих методов обнаружения.

Гибридная архитектура CNN-KAN представляет собой синтез двух методологически различных подходов: извлечения пространственных признаков с помощью сверточных нейронных сетей (CNN) и их последующей интерпретации с использованием сетей Колмогорова-Арнольда (KAN). Такое сочетание обеспечивает не только высокую точность при анализе изображений, но и интерпретируемость.

Модель CNN-KAN состоит из двух последовательных компонентов, выполняющих взаимодополняющие функции рисунок 1:

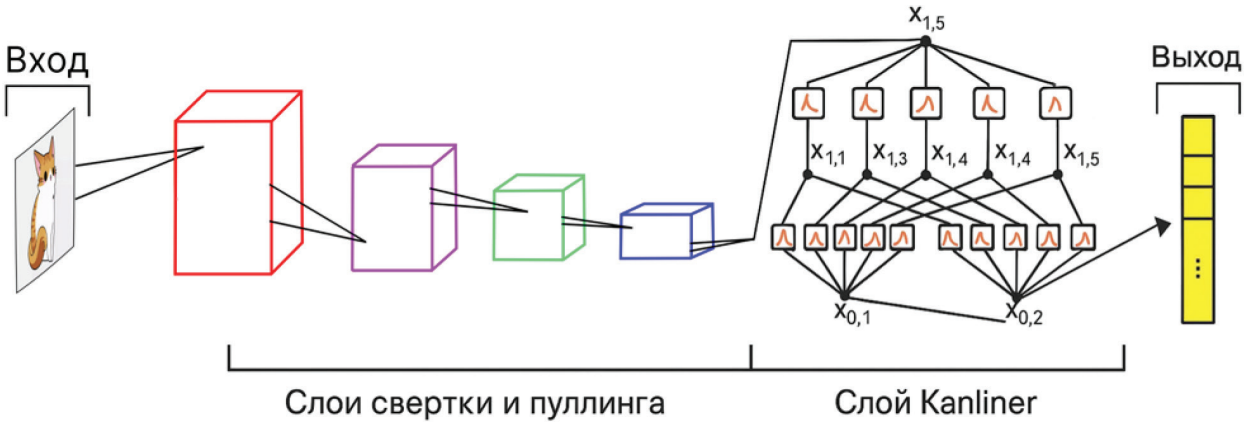

Рисунок 1. Схематичное представление архитектуры CNN-KAN

Сначала изображение обрабатывается сверточной нейросетью, которая отвечает за автоматическое извлечение локальных и глобальных признаков. Входной тензор в данном случае RGB-изображение размером 128×128 пикселя, последовательно проходит через набор сверточных фильтров — небольших матриц, «сканирующих» изображение с заданным шагом. Каждый фильтр нацелен на выделение определенного типа признаков: краев, текстур, цветовых переходов. За сверткой следует нелинейная активация, в данном случае SELU, которая устраняет отрицательные значения и усиливает выраженные паттерны. В отличие от более простых функций, таких как ReLU, SELU не только вносит нелинейность, но и помогает стабилизировать процесс обучения. Далее применяется операция пулинга, в данном случае MaxPooling, позволяющая сократить размер карты признаков и повысить устойчивость модели к искажениям. Например, окно 2×2 со сдвигом 2 сокращает пространственные размеры в два раза, при этом сохраняются наиболее выраженные характеристики.

Современные архитектуры, такие как ResNet и EfficientNet, строятся из иерархии сверточных блоков, в которых каждый последующий слой обобщает признаки, выявленные предыдущими. Начальные уровни определяют простые элементы, такие как линии и углы, средние выявляют текстурные структуры, а финальные формируют представления об объектах высокого уровня — например, глазах, колесах или клеточных ядрах. Особенность ResNet заключается в использовании остаточных связей, которые позволяют избежать проблемы исчезающих градиентов и сохраняют поток информации от начальных слоев к глубинным.

После того как CNN завершает извлечение признаков, полученные карты преобразуются в одномерный вектор с помощью операции Flatten или Global Average Pooling.

На этом этапе начинается работа блока KAN, который заменяет традиционный многослойный перцептрон. В отличие от MLP, использующего фиксированные функции активации, в KAN каждый входной признак обрабатывается через адаптивную одномерную функцию, параметры которой подбираются в процессе обучения рисунок 2.

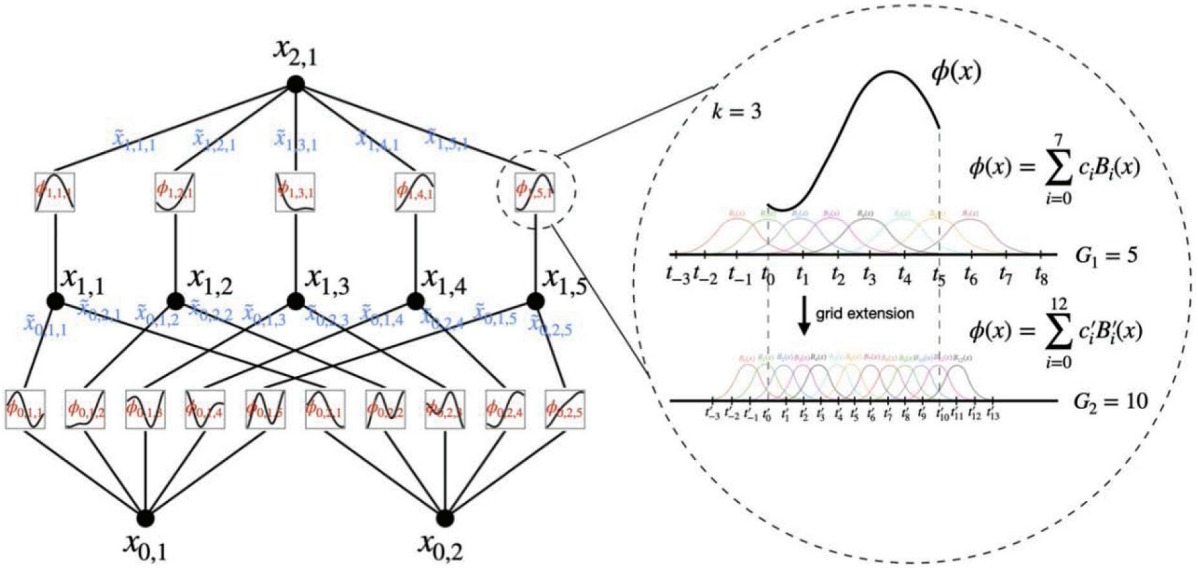

Рисунок 2. Схематичное представление KAN

Основная идея KAN основана на теореме Колмогорова-Арнольда, утверждающей, что любую многомерную непрерывную функцию можно представить как комбинацию одномерных функций и операций сложения. Это делает структуру модели интерпретируемой: каждое соединение в сети — это не просто вес, а аналитическая функция, которая может быть визуализирована и проанализирована отдельно.

Каждый признак, извлеченный CNN, пропускается через параметризованную функцию кубический сплайн и на выходе формируется значение, отражающее вклад этого признака в финальное предсказание. Все функции суммируются, создавая итоговую оценку модели. Таким образом, модель позволяет наглядно увидеть, как изменение одного признака влияет на результат.

Результаты сравнительного анализа представлены в таблице 1.

Таблица 1. Сравнение основных метрик качества всех исследуемых моделей и датасетов

Модель	Датасет	Accuracy (%)	Precision (%)	Recall (%)	F1-score	ROC-AUC
CNN-MLP	140k Real and Fake Faces	94,8	93,3	96,5	0,949	0,988
CNN-MLP	Deepfake and Real Images	87,8	87,8	87,5	0,876	0,948
CNN-KAN	140k Real and Fake Faces	96,8	96,9	96,7	0,968	0,994
CNN-KAN	Deepfake and Real Images	90,2	89,1	91,7	0,903	0,967

Из таблицы очевидно, что гибридная модель CNN-KAN стабильно демонстрирует более высокие показатели по всем ключевым метрикам по сравнению с моделью CNN-MLP на обоих датасетах. Это связано с уникальной архитектурой CNN-KAN, включающей слои KANLinear, способные адаптивно настраивать параметры активации, эффективно улавливая сложные нелинейные признаки в данных.

Матрицы ошибок, полученные в ходе экспериментов рисунки 3-4, также подтверждают общую тенденцию к более точной классификации со стороны модели CNN-KAN. Количество ложноположительных и ложноотрицательных результатов в CNN-KAN меньше по сравнению с CNN-MLP, что подтверждает практическую пригодность и надежность CNN-KAN для применения в реальных условиях.

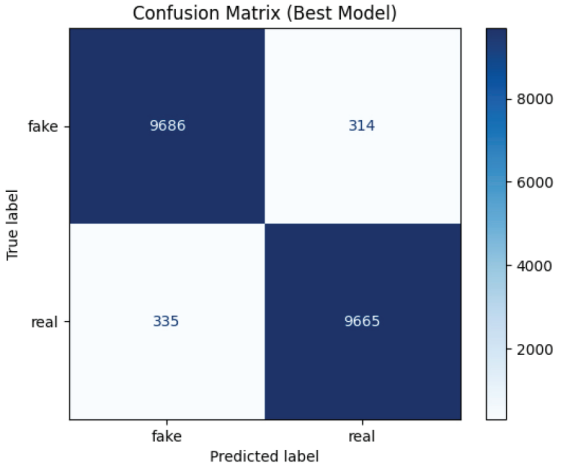

Рисунок 3. Матрица ошибок CNN-KAN

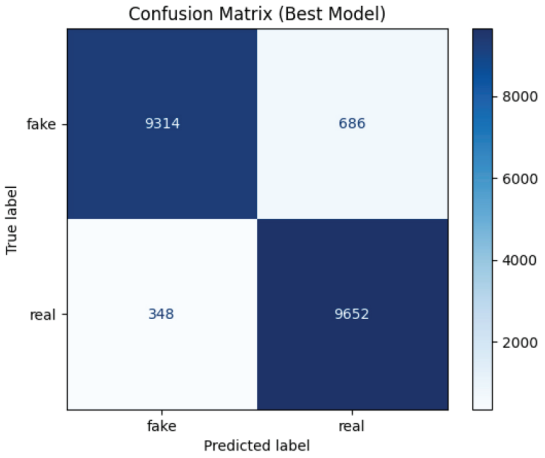

Рисунок 4. Матрица ошибок CNN-MLP

Архитектурные преимущества CNN-KAN, такие как адаптивные нелинейности и возможность динамического подстраивания архитектуры под конкретные задачи, способствуют ее более стабильной работе и высокой точности. Данный подход имеет большую перспективу для широкого применения в компьютерном зрении и других смежных областях, где требуется высокая точность и гибкость моделей.

Библиографический список

1. Адылова Ф.Т. Идея, основные разработки и применение нейронных сетей Колмогорова-Арнольда: аналитический обзор // Raqamli iqtisodiyot (Цифровая экономика). 2025. № 10. С. 7–15.

2. Воронин А.И., Гавра Д.П. Дипфейки: современное понимание, подходы к определению, характеристики, проблемы и перспективы // Российская школа связей с общественностью. 2024. № 33. С. 10–16.

3. Generative Adversarial Networks / Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio — Canada : Departement d'informatique et de recherche op ´ erationnelle ´ Universite de Montr ´ eal ´ Montreal, QC H3C 3J7 , 2014. 2-8 c. arXiv:1406.2661.

4. Deepfake Detection using Biological Features: A Survey / Kundan Patil, Shrushti Kale, Jaivanti Dhokey, Abhishek Gulhane — Mumbai : University at Albany, State University of New York, 2023. 2-13 c. arXiv:2301.05819v1.

5. KAN: Kolmogorov–Arnold Networks / Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, Max Tegmark — USA : Massachusetts Institute of Technology, California Institute of Technology, Northeastern University, The NSF Institute for Artificial Intelligence and Fundamental Interactions, 2024. 2-50 c. arXiv:2404.19756v4.

6. Kolmogorov-arnold network for satellite image classification in remote sensing / Minjong Cheon //arXivLabs: experimental projects with community collaborators — USA, 2024. С. 2–10.

7. Github platform for development. Kolmogorov-Arnold Networks (KANs): сайт. 2024. URL: https://github.com/KindXiaoming/pykan?tab=readme-ov-file (дата обращения: 05.05.2025).

8. Lecun Y., Bengio Y., Hinton G. Deep learning // Nature. 2015. Vol. 521. P. 436–444.

9. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge: MIT Press, 2016. 655 p.

10. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition // arXiv preprint arXiv:1409.1556. 2014. 13 p.

11. Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems. 2012. Vol. 25. P. 1097–1105.

12. Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift // International Conference on Machine Learning (ICML). 2015. P. 448–456.

13. Энциклопедия «Касперского». Дипфейк, deepfake: сайт. 2025. URL: https://encyclopedia.kaspersky.ru/glossary/deepfake/ (дата обращения: 10.05.2025).

14. Google DeepMind. SynthID: сайт. 2025. URL: https://deepmind.google/science/synthid/ (дата обращения: 10.05.2025).

15. Метрики качества моделей бинарной классификации: сайт. 2023. URL: https://loginom.ru/blog/classification-quality (дата обращения: 10.05.2025).