УДК 004.932.2

МЕТОД ИНТЕРПРЕТАЦИИ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ С ПОМОЩЬЮ ГРАФОВОГО ВНИМАНИЯ

Пикуль А.С. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Попов И.Ю. (ИТМО)

Введение. Дипфейки – поддельные видео или изображения, созданные на основе технологий искусственного интеллекта. С их помощью можно нанести значительный ущерб в разных социальных областях: распространение дезинформации, разрушение репутации и подрыв доверия к цифровым медиа или конкретной личности. В основе современных методов обнаружения дипфейков используют глубокие нейронные сети, которые работают как "черные ящики", что затрудняет их использование в критически важных сферах.

Основная часть. Для того чтобы объяснить работу черного ящика существуют классические методы интерпретации. К ним относят LIME [1], SHAP [2] и градиентные методы [3]. Данные методы построены на основе визуальной интерпретации. На данный момент наиболее известным методом является GradCam [4], который позволяет получить карты признаков, отражающих информацию о расположении объектов на исходном изображении.

Отдельным направлением развития методов интерпретации можно выделить методы на основе механизмов внимания. Основная идея таких подходов заключается в том, чтобы определить, какие части входных данных оказывают наибольшее влияние на предсказание модели, и показать их с помощью карт внимания.

В данной работе за основу был выбран механизм внимания, а именно, его модификация на основе графов [5]. Идея состоит в том, чтобы полученные из входного изображения карты характеристик разбивать на патчи. Теперь данные патчи будут исполнять роль узлов в графе. Далее между узлами графа вычисляются коэффициенты внимания. Коэффициенты показывают уровень связей между узлами: чем выше коэффициент, тем больше связь. Далее коэффициенты используются для построения карт внимания и последующей интерпретации работы модели.

Выводы. базовая модель с GATv1 продемонстрировала высокие результаты на наборе данных FF++ [6]: F1-score и AUC порядка 99,9%, что говорит о высокой способности модели извлекать релевантные признаки. Однако на наборе данных DFDC [7] модель показала более низкие метрики, около 80% F1-score и 85-86% AUC. Важно отметить, что переход от GATv1 к GATv2 принёс улучшение качества, что означает, что внедрение динамического внимания позволяет повысить точность детекции.

Эксперименты с разной глубиной GAT показали эволюцию карт внимания: от диффузного распределения фокуса по всему изображению на ранних слоях к более выделению ключевых областей лица на более глубоких слоях. Это говорит о том, что увеличение числа слоёв GAT способствует точной локализации важных областей.

Дополнительный эксперимент, в котором карта внимания объединялась с исходным тензором, продемонстрировал ещё больший прирост качества на DFDC. Этот подход сохранил в итоговом представлении информацию об исходных деталях, которую механизм внимания мог частично утратить, и одновременно подчеркнул важные признаки. В результате улучшение F1-score (85,29 %) и AUC (89,33 %) стало более существенным, что свидетельствует о необходимости комплексного учёта как локальных признаков, выделяемых вниманием, так и глобального контекста исходных данных.

На основе полученных карт внимания можно оценивать работу модели. По картам видно, что модель формирует итоговое решение на основе ключевых областей лица – глаза, брови, нос, щеки, подбородок, лоб, уши. Так как именно эти области являются ключевыми для

дипфейков, можно сделать вывод о том, что модель верно определяет ключевые области для детекции. Также видно, что классические методы справляются с определением ключевых областей лица хуже, чем разработанный подход.

Список использованных источников:

1. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, pp. 1135–1144, 2016.
2. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.
3. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. 34th Int. Conf. Machine Learning, pp. 3319–3328, 2017.
4. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision, pp. 618–626, 2017.
5. Petar Velickovic. Graph Attention Networks / Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio // At ICLR. – 2018.
6. Rossler A. et al. Faceforensics++: Learning to detect manipulated facial images //Proceedings of the IEEE/CVF international conference on computer vision. – 2019. – С. 1-11.
7. Dolhansky B. The dee pfake detection challenge (DFDC) pre view dataset //arXiv preprint arXiv:1910.08854. – 2019.