Scientific article
УДК 004.056
DOI: 10.26583/bit.2025.2.06

ENHANCING EXPLAINABILITY IN DEEPFAKE DETECTION WITH GRAPH ATTENTION NETWORKS

Aleksandr S. Pikul[1], Ilya Yu. Popov[2]
National Research University ITMO, St. Petersburg, 197101, Russia
[1]aleksandrpikul134@gmail.com, https://orcid.org/0009-0000-2619-3822
[2]ilyapopov27@gmail.com, https://orcid.org/0000-0002-6407-7934

Abstract. Understanding how artificial intelligence models make decisions is important, especially for difficult tasks like detecting deepfakes, where it's not enough to just get a result – it needs to know why the model made that choice. Many current methods, like Shapley additive explanations (SHAP) and Gradient-weighted Class Activation Mapping (Grad-CAM), help explain these decisions, but they often aren't detailed enough for tasks involving complex data like human faces. In this paper, it's introduces a new method that uses Graph Attention Networks (GATs) to explain deepfake detection. Instead of looking at images as a whole, it's turn the face into a graph, where each key part of the face (like the eyes, nose, and mouth) is a separate node. This helps the model focus on the most important areas. Using attention mechanisms, the model highlights which parts of the face influenced its decision, making the process easier to understand. It's compares two versions of the model, GATv1 and GATv2, and show how both create clear visual explanations while still performing well in detecting deepfakes. This approach makes it easier to see how the model reaches its conclusions, improving trust and transparency. The code is freely available at https://github.com/aleksandrpikul/ResGAT/tree/main.

Keywords: *deepfake, explainability, attention, deepfake detection, graph attention network.*

For citation: *Pikul, Aleksandr S.; Popov, Ilya Yu. Enhancing explainability in deepfake detection with graph attention networks. IT Security (Russia), [S.l.], v. 32, no. 2, p. 73–82, 2025. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1776. DOI: http://dx.doi.org/10.26583/bit.2025.2.06.*

****Acknowledgement.*** This research was supported by ITMO University (Project No. 623106 “Autonomous intelligent systems”).

Научная статья

ПОВЫШЕНИЕ ОБЪЯСНИМОСТИ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ С ПОМОЩЬЮ ГРАФОВОЙ СЕТИ НА ОСНОВЕ СОСРЕДОТОЧЕНИЯ ВНИМАНИЯ

Александр С. Пикуль[1], Илья Ю. Попов[2]
Университет ИТМО, Кронверкский пр., 49, Санкт-Петербург, 197101, Россия
[1]aleksandrpikul134@gmail.com, https://orcid.org/0009-0000-2619-3822
[2]ilyapopov27@gmail.com, https://orcid.org/0000-0002-6407-7934

Аннотация. Понимание о том, как именно модели искусственного интеллекта принимают решения, важно, в особенности для таких сложных задач, как обнаружение дипфейков, где недостаточно просто получить результат – необходимо знать, почему именно модель приняла такое решение. Многие существующие методы интерпретации моделей искусственного интеллекта, такие как SHAP и Grad-CAM, помогают объяснить полученные решения, но зачастую полученные объяснения недостаточно детализированы для таких сложных данных как человеческие лица и дипфейки. В данной статье представлен новый метод с использованием графовых сетей с механизмом внимания (GAT) для задачи интерпретации решений детекторов дипфейков. Вместо того чтобы рассматривать изображение в целом, оно разбивается на патчи, а затем формируются в граф, где каждая ключевая часть лица (глаза, нос и рот) является отдельным узлом. Такое разбиение помогает модели учитывать при принятии решений наиболее важные области изображения. С помощью механизма внимания, модель выявляет, какие области изображения и лица повлияли на её решение. Проведено сравнивние двух версий механизма внимания, GATv1 и GATv2, и показано, как создается визуальная интерпретация с их помощью. При этом сохраняется высокая точность для классификатора дипфейков. Предложенный подход повышает понимание о том, как модель обрабатывает изображения, то есть, какие признаки оказались наиболее важными при принятии решения. Таким образом повышается доверие к работе модели искусственного интеллекта. Код в открытом доступе по адресу https://github.com/aleksandrpikul/ResGAT/tree/main.

Ключевые слова: *дипфейки, объяснимость, механизм внимания, обнаружение дипфейков, графовая сеть с вниманием.*

Для цитирования: *Пикуль, Александр С.; Попов, Илья Ю. Повышение объяснимости обнаружения дипфейков с помощью графовой сети на основе сосредоточения внимания. Безопасность информационных технологий, [S.l.], т. 32, № 2, с. 73–82, 2025. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1776. DOI: http://dx.doi.org/10.26583/bit.2025.2.06.*

Introduction

Deepfakes are highly realistic fake videos or images created using artificial intelligence, making individuals appear to say or do things they never actually did. These manipulations pose significant threats by spreading misinformation, damaging reputations, and undermining trust in digital media. As a result, detecting deepfakes is crucial to mitigate their negative impacts and ensure the integrity of information across various platforms.

Current methods for detecting deepfakes primarily rely on deep learning models, which are powerful tools capable of identifying subtle alterations in media. However, these models often operate as "black boxes," providing results without any explanation of how they arrived at their conclusions. This lack of transparency poses challenges, especially in critical areas such as justice, medicine, and security, where understanding the reasoning behind a decision is as important as the decision itself.

To address the opacity of black-box models, several classical interpretation methods have been developed. These techniques aim to shed light on the inner workings of deep learning models by providing insights into their decision-making processes. Some of the most widely used classical interpretation methods include:

Feature Importance Scores. This method assesses the contribution of each input feature to the model's prediction. By ranking features based on their importance, users can identify which aspects of the data most influence the outcome. Techniques like permutation importance and mean decrease in impurity fall under this category [1]. However, feature importance scores can sometimes be misleading, as they may not capture complex interactions between features.

LIME (Local Interpretable Model-Agnostic Explanations). LIME [2] approximates the behavior of a black-box model locally around a specific prediction by fitting a simple, interpretable model, such as a linear regression or decision tree. This allows users to understand the factors driving individual predictions. While LIME is versatile and can be applied to various types of data, its explanations are limited to local regions and may not reflect the model's global behavior.

Grad-CAM. Grad-CAM [3] is a visualization method used to interpret the decisions of deep neural networks, particularly in computer vision tasks. It leverages the gradients flowing into the final convolutional layer of the network to highlight important regions in the input image that influenced the model's prediction. Grad-CAM generates activation maps that show which parts of the image the model focuses on when making a decision. For example, in image classification tasks, Grad-CAM allows for visual inspection of the objects or regions that the network considers crucial for its classification.

SHAP (SHapley Additive exPlanations). SHAP [4] assigns each feature an importance value based on cooperative game theory, ensuring that the explanations are consistent and theoretically sound. SHAP values provide a unified measure of feature contribution across different models. However, computing SHAP values can be computationally intensive, especially for large and complex models.

Partial Dependence Plots (PDPs). PDPs [5] show the relationship between a subset of input features and the predicted outcome, averaging out the effects of all other features. This helps in understanding the marginal effect of certain features on the prediction. While PDPs provide a global view, they can oversimplify interactions between features.

Integrated Gradients. This method attributes the prediction of a deep network to its input features by integrating gradients along a path from a baseline input to the actual input [6]. Integrated Gradients help in understanding the contribution of each feature in a way that satisfies certain axioms, ensuring more reliable explanations.

While these classical interpretation methods offer valuable insights, they often provide limited explanations or fail to capture the intricate relationships within the data, particularly in the context of deepfake detection. For instance, many of these methods focus on identifying important features in isolation rather than understanding the complex interactions and dependencies between multiple features that may contribute to a deepfake.

This gap underscores the need for more sophisticated approaches that can offer deeper and more meaningful explanations. In this article, we aim to enhance the explainability of deepfake detectors by introducing a novel approach based on graph neural networks with attention mechanisms. Unlike traditional methods, our approach models the relationships between different features in the data, allowing for a more comprehensive understanding of how decisions are made. The attention mechanism further highlights the most important elements and their interactions, providing clearer and more detailed explanations. By improving transparency beyond what classical interpretation methods offer, we seek to build greater trust and effectiveness in deepfake detection systems, especially in vital areas where understanding the "why" behind a decision is crucial.

1. Interpretation methods for deepfake detection

This chapter reviews various methods used to interpret deepfake detection models, highlighting both classical and modern approaches. We examine different techniques based on a range of studies and identify key features that enhance the explainability of these models.

One of the most detailed studies in this area is the master's thesis presented in [7]. The primary goal of this work is to improve the explainability of deepfake detectors by introducing three distinct approaches:

- Image Segmentation. This method involves identifying specific areas of an image that the model considers fake. By highlighting these regions, users can see which parts of the image are contributing to the model's decision.
- Multi-Class Classification. This approach classifies different attributes of the fake content, such as identifying fake facial features. This helps in understanding which specific features are being manipulated.
- Text Generation (Captioning). This technique generates natural language explanations that describe why a particular video or image is considered fake. These textual explanations make it easier for users to understand the model's reasoning.

In [8], a simple pipeline is presented that uses convolutional neural networks (CNNs) and capsule networks to detect deepfakes. Although this approach is simple and easy to understand, it focuses on a single variation of the pipeline. The study emphasizes the importance of combining

CNNs with Capsule Networks to improve detection accuracy and provide some level of interpretability.

In [9] the researchers propose a pipeline where image segmentation results are combined with heatmaps. The segmentation highlights fake regions, and the heatmap provides additional visual cues about the model's focus areas. These combined visualizations are then fed into the deepfake detector for final classification. This method not only interprets the model's decisions but also integrates multiple sources of information to enhance accuracy.

In [10] the authors introduce a comprehensive pipeline that includes several steps to improve explainability:

- Feature Extraction. Uses CNNs like ResNet to extract high-level features from each video frame.
- Feature Aggregation. Combines features from consecutive frames to capture temporal dependencies.
- Temporal Analysis. Employs Long Short-Term Memory (LSTM) networks to analyze changes over time, identifying inconsistencies typical of deepfakes.
- Transparency and Explainability. Utilizes Grad-CAM to visualize important areas in each frame and SHAP values to quantify the contribution of each feature to the model’s predictions.
- Classification. Finalizes the detection using a fully connected layer with a sigmoid activation function to classify videos as real or fake.
- Model Evaluation. Assesses the model using metrics like accuracy, recall, precision, F1-score, and ROC AUC.
- User Studies. Gathers feedback from users to evaluate the clarity of visualizations and explanations provided by Grad-CAM and SHAP.

This multi-step approach not only enhances detection performance but also provides detailed explanations, making the model’s decisions more transparent.

In [11] the authots explore the use of attention mechanisms for interpretability. This method involves creating a hierarchical ensemble of models that focus on different aspects of the data. The attention mechanism helps in highlighting the most relevant features and their interactions, providing a clear explanation of why a particular decision was made.

In [12] the authors combine classical interpretation methods like SHAP and LIME with newer techniques such as attention mechanisms. This hybrid approach leverages the strengths of both types of methods to provide more accurate and understandable explanations. Additionally, adversarial robustness training is used to make the model more resilient to attempts to deceive it.

In [13] the authors present an implementation of an attention-based deepfake detection model. By incorporating attention mechanisms, the model can highlight which parts of the input data are most influential in making its predictions. This not only improves detection performance but also provides clear visual explanations for users.

The reviewed approaches share several key features that enhance the explainability of deepfake detection models:

1. Multimodality. Interpretations are provided not only through visual means, such as heatmaps and segmentation, but also through textual explanations. This makes the explanations more comprehensive and easier to understand for users.

2. Combination of Classical and Modern Methods. Alongside traditional methods like SHAP and Grad-CAM, newer techniques such as attention mechanisms are employed. This combination leverages the strengths of both types of methods to provide more detailed and reliable explanations.

3. Integration of Multiple Approaches. For more accurate interpretations, some methods combine several techniques. For example, feature maps may be combined with attention maps to give a fuller picture of what the model is focusing on.

4. Application in Classification. Interpretative methods are often integrated directly into the classification process. For instance, feature maps generated from segmentation can be combined with heatmaps and passed to the classifier, enhancing both the detection accuracy and the interpretability of the model.

These features collectively contribute to more transparent and trustworthy deepfake detection systems, making it easier for users to understand and trust the model's decisions.

2. Graph neural networks for explainable deepfake detection

While traditional interpretation methods provide valuable insights into deepfake detection models, there are emerging approaches that leverage Graph Neural Networks (GNNs) and graph attention mechanisms to enhance explainability. These methods offer a different perspective by modeling relationships and interactions between various features, providing more comprehensive and intuitive explanations.

Graph Neural Networks (GNNs) [14] are a class of neural networks designed to work with graph-structured data. Unlike traditional neural networks that operate on fixed-size input data, GNNs can effectively capture the relationships and interactions between nodes (entities) in a graph. This capability makes GNNs particularly suited for tasks where the data is naturally represented as a graph, such as social networks, molecular structures, and, increasingly, in explainable AI for deepfake detection.

Explainable AI (XAI) aims to make the decision-making processes of complex models understandable to humans. By integrating GNNs into deepfake detection systems, researchers can create models that not only identify fake content but also provide insights into how different features interact to influence the final decision.

Several studies have explored the use of GNNs to enhance the explainability of deepfake detection models. These approaches typically involve representing video frames or image features as nodes in a graph, with edges capturing the relationships between these features. By processing this graph structure, GNNs can learn to identify patterns and inconsistencies that are indicative of deepfakes.

Graph Attention Networks (GATs) extend GNNs by incorporating attention mechanisms, which allow the model to weigh the importance of different nodes and edges dynamically. This feature is particularly useful for explainability, as it enables the model to focus on the most relevant parts of the graph when making a prediction.

In [15] the authors introduce attention-based techniques that highlight critical nodes and edges contributing to the model's decisions. This approach not only improves the interpretability of the model but also enhances its ability to generalize across different types of deepfake content.

In [16] the authors employ hierarchical GNNs with attention mechanisms to analyze deepfake videos. The attention layers help in identifying key features and their interactions, providing clear visual explanations of why a video is classified as fake.

3. Proposed method for explainable deepfake detection

The proposed architecture (Fig. 1), as shown in the figure, begins by passing an image through a ResNet encoder for feature extraction. These extracted features are then transformed into graph data and passed through a Graph Attention Network (GAT). The attention mechanism in the GAT highlights the most important nodes in the graph, providing a way to visualize which parts of the image contributed most to the model's decision. This architecture is modified based on [17].

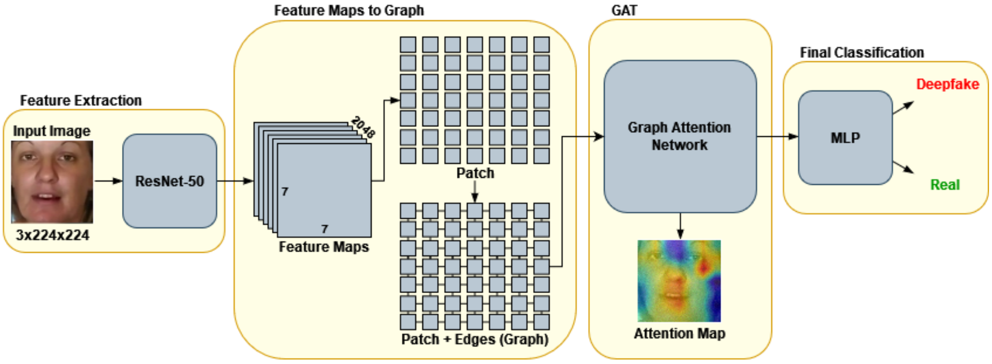

Fig. 1. The proposed architecture (ResGAT)

Feature Extraction. The ResNet-50 [18] model is used as the backbone for feature extraction. This pre-trained network has been widely adopted in image classification tasks and is capable of extracting rich representations from the input image. The last fully connected layer of ResNet is removed, and the output is a feature map of size (batch_size, 2048, 7, 7), where 7x7 represents the spatial resolution and 2048 is the number of feature channels.

Graph Construction. The extracted feature map is reshaped into a graph where each spatial location (e.g., a 7x7 grid) corresponds to a node in the graph. We create edges between neighboring nodes to represent spatial relationships, ensuring that each node has connections to its right and bottom neighbors.

For each image, this results in a graph with 49 nodes (corresponding to the 7x7 grid), and each node is connected to its neighboring nodes in a structured grid layout. The edges capture the relationships between adjacent pixels or regions.

Graph Attention Network (GAT). The core component of the architecture is a Graph Attention Network (GAT) that processes graph-structured data. GAT assigns different attention weights to the edges, allowing the model to focus on the most relevant nodes and their connections. This helps the model highlight the most significant regions of the image, which improves the detection of deepfakes.

We use two versions of GAT in our architecture. The first version, GATv1 [19], computes attention for each node towards its neighbors based on their features. GATv1 uses a static attention mechanism, meaning that the ranking (order) of attention for key nodes remains the same for all queries. In other words, for any query node, the ranking of attention to its neighboring nodes stays static, which can limit the model's flexibility. The formula for GATv1 is as follows:

$$e_{ij} = LeakyReLU(a^T[Wh_i||Wh_j]), \tag{1}$$

where:

- e_{ij} is the attention value for node i towards node j,
- W is a trainable weight matrix for transforming node features,
- h_i and h_j are the features of nodes i and j,
- a is the attention vector.

The second version, GATv2 [20], addresses the limitations of static attention in GATv1. GATv2 introduces dynamic attention, where the attention weights for neighboring nodes are computed dynamically for each query node. This allows the model to adapt more flexibly to different data and situations. The formula for GATv2 is as follows:

$$e_{ij} = a^T LeakyReLU([Wh_i||Wh_j]). \tag{2}$$

In GATv2, the attention weights for each node are dynamically calculated based on both the node's own state and the state of its neighbors. This provides greater flexibility in how the model attends to different regions of the image.

Multi-Layer Perceptron (MLP). After processing the graph through the GAT layer, we pool the node features by averaging them and passing them through a fully connected layer (MLP). This layer is responsible for classifying the input as either a deepfake or a real image. The output is a probability score that indicates the likelihood of the input being a deepfake.

Attention Visualization. One of the key features of our architecture is the ability to visualize the attention weights generated by the GAT layer. These attention weights highlight which nodes were most influential in the model's decision. We create attention maps that show these important regions.

4. Experiments

The experiments were conducted in two stages to evaluate the effectiveness of the proposed model and compare the performance of two versions of the GAT: GATv1 and GATv2.

Datasets:

- DFDC [21]. The DFDC dataset is one of the largest and most diverse collections of deepfake videos. It includes real and manipulated videos of people speaking, with various types of video tampering techniques applied. In this experiment, frames were extracted from the videos, and the following augmentations were applied: blurring, noise, transposition, rotation, and resizing. Training set: 32,300 deepfakes and 47,700 real faces. Testing set: 4,027 deepfakes and 4,639 real faces.
- FF++ [22]. The FF++ dataset contains both real and manipulated videos, created using various deepfake generation methods, including face swaps and neural networks. Unlike the DFDC dataset, no augmentations were applied to the extracted frames in this experiment. Training set: 300,252 deepfakes and 220,496 real faces. Testing set: 75,063 deepfakes and 55,124 real faces.

Given the imbalance between real and fake faces in the DFDC and FF++ datasets, the F1-score was chosen as the primary evaluation metric. This metric is particularly useful in cases of class imbalance, as it takes into account both precision (the proportion of correctly identified deepfakes) and recall (the ability to find actual deepfakes). By using the F1-score, we get a more reliable measure of the model's performance, ensuring it handles both real and fake faces effectively, even when the classes are not evenly distributed.

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{3}$$

The experiments were conducted in two stages: first with GATv1, then with GATv2. The table 1 presents the F1-score and AUC values for two versions of the model (GATv1 and GATv2) on the DFDC and FF++ datasets.

Table 1. Results For GATv1 and GATv2

Method	F1-score, %		AUC	
	FF++	DFDC	FF++	DFDC
GATv1	99.68	80.01	99.9	85.92
GATv2	99.93	81.72	99.9	85.96

Fig. 2 and 3 present attention maps for all experiments, as well as visualizations using Grad-CAM and LIME for comparison with the proposed model.

In both datasets, the proposed version GATv2 demonstrates a higher F1-score, indicating improvements due to the dynamic attention mechanism. The difference in results between the two datasets is explained by their structure and the application of augmentations. The FF++ dataset contains more images and diverse forgeries, enabling the model to generalize its predictions better.

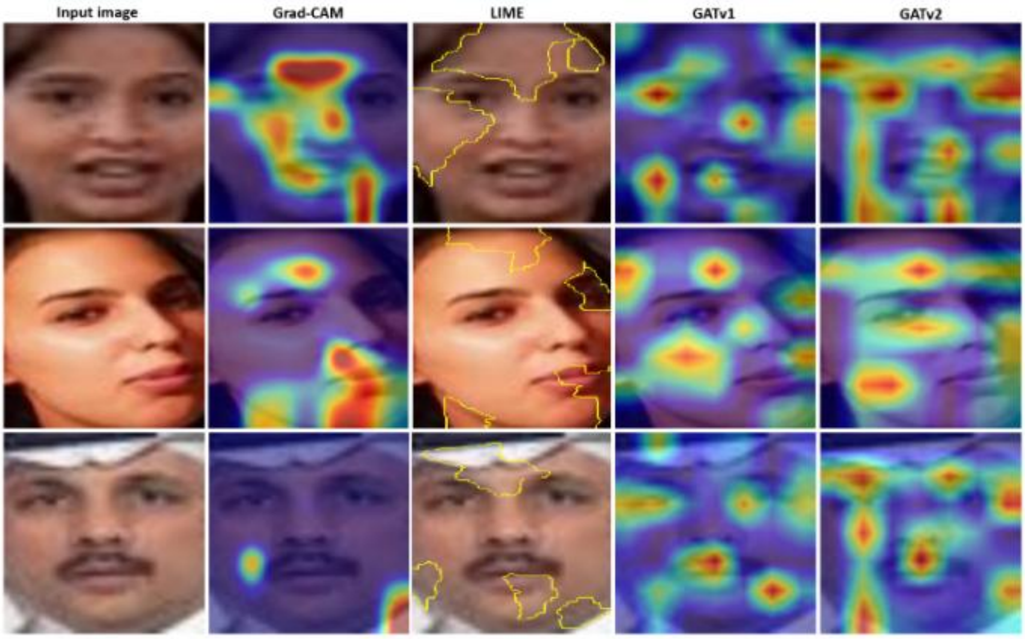

Fig. 2. Attention maps from GATv1 and GATv2 on FF++ (comparison with Grad-CAM and LIME)

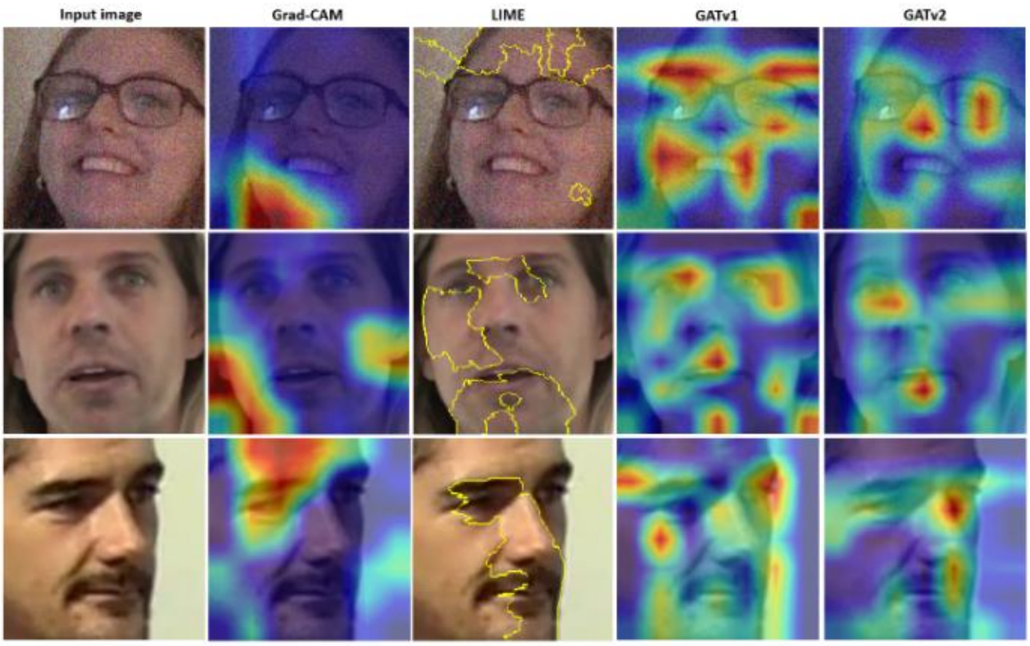

Fig. 3. Attention maps from GATv1 and GATv2 on DFDC (comparison with Grad-CAM and LIME)

From the visualizations, it is clear that the proposed model focuses more accurately on key areas of the face compared to Grad-CAM and LIME.

Fig. 4 demonstrates a comparison of attention maps for a real face and a deepfake. As can be seen from the examples, for both the real face and the deepfake, the model focuses its attention on key facial areas, including the eyes (along with their shape and contours), nose, mouth, chin, cheeks, and forehead.

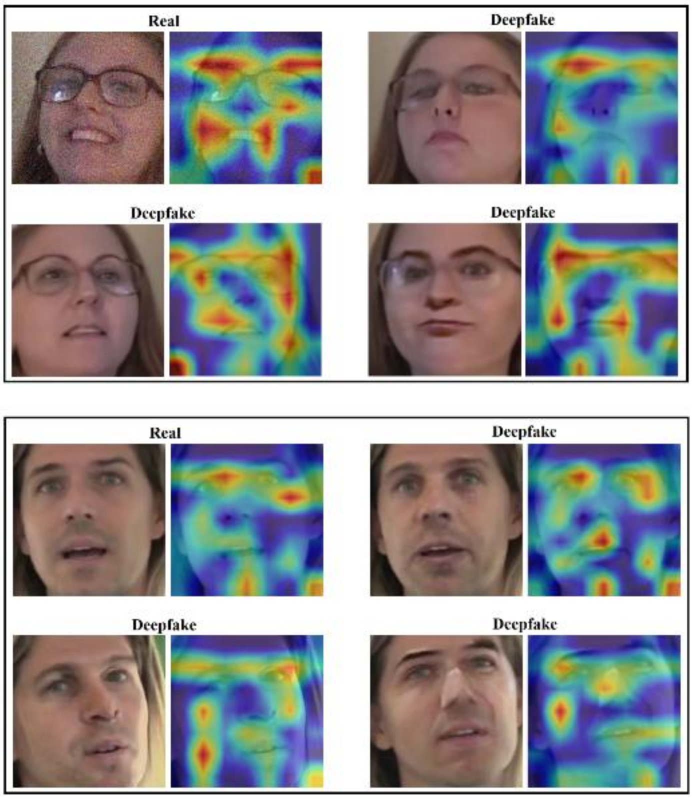

Fig. 4. Attention Maps for Deepfakes and Real Faces

Fig. 4 demonstrates a comparison of attention maps for a real face and a deepfake. As can be seen from the examples, for both the real face and the deepfake, the model focuses its attention on key facial areas, including the eyes (along with their shape and contours), nose, mouth, chin, cheeks, and forehead.

5. Conclusion

The experiments conducted on the DFDC and FF++ datasets demonstrated that our GAT-based approach not only improved performance, particularly with GATv2, but also offered better visualizations of attention maps compared to existing techniques like Grad-CAM and LIME. These attention maps allowed us to accurately interpret the areas that influenced the model's classification, providing both transparency and reliability in deepfake detection.

In experiments on the DFDC and FF++ datasets, we observed improvements in detection accuracy, especially with GATv2, which achieved an F1-score of 81.72% on the DFDC dataset and 99.93% on the FF++ dataset.

REFERENCES:

[1] Breiman L. Random Forests. Machine Learning, v. 45, no. 1, p. 5–32, 2001. DOI: http://dx.doi.org/10.1023/A:1010933404324.

[2] Ribeiro M.T., Singh S., and Guestrin C.Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, p. 1135–1144, 2016. DOI: 10.1145/2939672.2939778.

[3] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proc. IEEE Int. Conf. Computer Vision, p. 618–626, 2017. DOI: 10.1109/ICCV.2017.74.

[4] Lundberg S.M. and Lee S.-I. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst., v. 30, p. 4765–4774, 2017. DOI: https://doi.org/10.48550/arXiv.1705.07874.

[5] Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Statist., v. 29, no. 5, p. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[6] Sundararajan M., Taly A., and Yan Q. Axiomatic Attribution for Deep Networks. Proc. 34th Int. Conf. Machine Learning, p. 3319–3328, 2017. URL: http://proceedings.mlr.press/v70/sundararajan17a.html (accessed: 07.12.2025).

[7] Fernandez I. Sanchez. Towards Multi-Granular Explainable AI: Increasing the Explainability Level for Deepfake Detectors. M.S. thesis, Univ. Oulu, 2022. URL: https://www.politesi.polimi.it/handle/10589/180027 (accessed: 07.12.2025).

[8] Singh A., Agarwal S., Chauhan P. and Sinha A. Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet. Proc. Int. Conf. Machine Learning Applications, 2021. DOI: https://doi.org/10.48550/arXiv.2404.12841.

[9] Mathews S., Trivedi S., House A., Povolny S. and Fralick C. An explainable deepfake detection framework on a novel unconstrained dataset. J. Advances in Science and Research, v. 9, p. 4425–4437, Jan. 2023. URL: https://link.springer.com/article/10.1007/s40747-022-00956-7 (accessed: 07.12.2025).

[10] Guo D., Wang H., Wang Y., Liang J., Zhang J. and Chang M. DeepExplain: Enhancing DeepFake Detection Through Transparent and Explainable AI Model. J. Artif. Intell. Res., v. 70, p. 593–626, 2021. DOI:10.31449/inf.v48i8.5792.

[11] Zhou P., Han X., Morariu V.I., andDavis L.S. Two-Stream Neural Networks for Tampered Face Detection. Proc. IEEE Conf. Computer Vision Pattern Recognit. Workshops, p. 1831–1839, 2018. DOI: 10.1109/CVPRW.2017.229.

[12] Johnson M., Smith L. and Nguyen T. Securing Online Integrity: A Hybrid Approach to Deepfake Detection and Removal using Explainable AI and Adversarial Robustness Training. Proc. AAAI Conf. Artif. Intell., p. 1234–1241, 2021. DOI: https://doi.org/10.1080/00051144.2024.2400640.

[13] Chawla V. Explainable Attention Based Deepfake Detector. GitHub repository. URL: https://github.com/vineethChawla/Explainable_Attention_Based_Deepfake_Detector (accessed: 07.12.2025).

[14] Wu Z., Pan S., Chen F., Long G., Zhang C., and Philip S.Y. A Comprehensive Survey on Graph Neural Networks. IEEE Trans. Neural Netw. Learn. Syst., v. 32, no. 1, p. 4–24, 2020. DOI: 10.1109/TNNLS.2020.2978386.

[15] Pope P.E., Kolouri S., Rostami M., Martin C.E. and Hoffmann H. Explainability Methods for Graph Convolutional Neural Networks. Proc. IEEE/CVF Conf. Computer Vision Pattern Recognit., p. 10772–10781, 2019. DOI: 10.1109/CVPR.2019.01103.

[16] Silva S.H., Bethany M., Votto A.M. and Scarff I.H. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Science International Synergy, v. 4, no. 1, Jan. 2022. DOI: 10.1016/j.fsisyn.2022.100217.

[17] Pikul A.S. and Lependin A.A. Deep neural network with graph attention for detecting fake facial images. Conference paper, v. 20, p. 190–193, 2023. URL: https://www.elibrary.ru/item.asp?id=69182160 (accessed: 07.12.2024) (in Russian).

[18] He K., Zhang X., Ren S. and Sun J. Deep Residual Learning for Image Recognition. Proc. IEEE Conf. Computer Vision Pattern Recognit., p. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[19] Veličković P., Cucurull G., Casanova A., Romero A., Liò P. and Bengio Y. Graph Attention Networks. Int. Conf. Learn. Represent., 2018. URL: https://arxiv.org/abs/1710.10903 (accessed: 07.12.2025).

[20] Brody S., Alon U. and Yahav E. How Attentive are Graph Attention Networks?Int. Conf. Learn. Represent., 2021. URL: https://arxiv.org/abs/2105.14491 (accessed: 07.12.2025).

[21] Dolhansky B., Howes R., Pflaum B., Baram N. and C.C. Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv:2006.07397, 2019. URL: https://arxiv.org/abs/2006.07397 (accessed: 07.12.2024).

[22] Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J. and Nießner M. FaceForensics++: Learning to Detect Manipulated Facial Images. Proc. IEEE Int. Conf. Computer Vision, p. 1–11, 2019. DOI: 10.48550/arXiv.1901.08971.

Статья поступила в редакцию 07.12.2024; одобрена после рецензирования 27.02.2025; принята к публикации 15.05.2025

The article was submitted 07.12.2024; approved after reviewing 27.02.2025; accepted for publication 15.05.2025