International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Emerging Trends and Innovations in Web-Based Applications and Technologies
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Deepfake Detection with Generalization via Domain-Aware Meta-Learning

Kartik. S. Raut.[1], Harsh. S. Pendke[2], Prof. Anupam Chaube[3], Prof. Usha Kosarkar[4]

[1,2,3,4]Department of Science and Technology,
[1,2]G H Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India
[3,4]G H Raisoni College of Engineering and Management, Nagpur, Maharashtra, India

ABSTRACT

Deepfakes, media manipulated using deep learning techniques, pose a growing threat to the integrity of digital content. These AI-generated forgeries are becoming increasingly sophisticated, making them difficult to detect. Traditional detection methods often lag behind the rapid evolution of deepfake techniques and are hampered by the limited variety of training data, making it hard for them to generalize effectively to new types of deepfakes. This thesis introduces a novel deepfake detection approach that combines meta-learning for domain generalization (MLDG) with self blended images (SBI) to address this challenge. MLDG, inspired by meta-learning principles, aims to improve the model's adaptability to new manipulation techniques by simulating domain shifts during training. The model learns from various source domains representing different deepfake generation methods. Additionally, SBIs, synthetic images created by blending real and manipulated faces, are incorporated to further diversify the training data and promote the learning of features that generalize across domains. This thesis focuses on detecting image-based deepfakes using the Face Forensics++ dataset, a benchmark collection of real and manipulated videos, specifically designed for deepfake detection research. The proposed method is evaluated with a leave-one-out cross-validation scheme on this dataset, where each deepfake generation technique is used as a test case while the others are used for training. The results consistently show that MLDG, when enhanced with SBIs, outperforms the standard Empirical Risk Minimization (ERM) method, demonstrating its effectiveness in generalizing to unseen manipulation techniques. The research offers a practical solution for deepfake detection, highlighting how MLDG and SBI augmentation can create more effective and adaptable detection systems. The findings emphasize the need for models that can adapt to evolving deepfake techniques to protect the integrity of digital media.

I. INTRODUCTION

The rapid advancements in deepfake generation technology have introduced significant challenges for detection methods, highlighting the need for more robust and adaptable approaches. Traditional deepfake detection relies on machine learning models trained to distinguish real and manipulated media based on specific deepfake techniques. However, these models often struggle when encountering previously unseen manipulation methods (Malik et al., 2022). Recent research efforts have aimed to improve the adaptability of detection models across different types of deepfake manipulations (Sun et al., 2021; Z. Wang et al., 2023), yet there remains a notable gap in model performance against the latest deepfake techniques, which can evade current detection systems (Le et al., 2024).

One promising approach to address the limitations in generalizing deepfake detection involves utilizing a **meta-learning framework for domain generalization**, combined with **data augmentation**. Meta-learning enhances adaptability by exposing the model to a variety of manipulations during training, enabling it to learn features that remain consistent across different deepfake generation techniques. Meanwhile, data augmentation increases the diversity of training data, further strengthening the model's ability to detect unseen deepfakes.

This study evaluates the effectiveness of an **integrated deepfake detection framework** that combines **Meta-Learning for Domain Generalization (MLDG) (D. Li et al., 2017b)** with **Self-Blended Images (SBIs) (Shiohara & Yamasaki, 2022)** as a data augmentation strategy. The primary objective is to assess how well this approach enhances detection performance on previously unseen deepfake generation techniques compared to a conventional baseline.

The evaluation will be conducted using the **FaceForensics++ dataset (Rössler et al., 2019)** through a **leave-one-out cross-validation** methodology. The model will be trained on multiple deepfake generation techniques and tested on unseen methods. The performance of the proposed approach will be compared against a **baseline model trained using Empirical Risk Minimization (ERM) (Vapnik, 1999)** to determine its effectiveness.

This study focuses exclusively on **video-based deepfake detection**, specifically targeting facial manipulations. While audio-based deepfake detection is an emerging concern, it is beyond the scope of this research. Additionally, the detection strategy used in this study is **image-based**, analyzing individual video frames instead of incorporating temporal information. This allows for a focused evaluation of domain generalization techniques applied to facial deepfake detection.

II. RELATED WORK

Challenges in Traditional Machine Learning for Deepfake Detection

Conventional machine learning models face several limitations, particularly their dependence on large labeled datasets for effective learning. In real-world deepfake detection, obtaining labeled datasets for every possible manipulation technique is challenging, time-consuming, and often impractical (Prince, 2023). Additionally, machine learning models frequently struggle with adapting to evolving or unseen tasks. For example, facial recognition

systems must adjust to new conditions, such as identifying individuals wearing masks (Batagelj et al., 2021), and spam filters must continuously adapt to detect novel phishing techniques (Alhogail & Alsabih, 2021). These challenges highlight the need for models that can generalize across different variations of deepfake manipulations without requiring extensive retraining.

III. Meta-Learning and Domain Generalization

Meta-Learning Frameworks

Meta-learning aims to improve a model's ability to learn new tasks efficiently by leveraging knowledge from previously encountered tasks. **Finn et al. (2017)** introduced a **task-based formalization** of meta-learning, accommodating both supervised and reinforcement learning settings. The **task-distribution view** conceptualizes meta-learning as optimizing **meta-knowledge (ω)**, which captures cumulative insights from multiple tasks. This enables the model to adapt quickly to unseen tasks while avoiding overfitting to training-specific features (Hospedales et al., 2020).

Meta-learning algorithms can be categorized into:

- **Model-based approaches** that leverage neural networks, such as **Recurrent Neural Networks (RNNs) or transformers**, to process task information efficiently (Munkhdalai & Yu, 2017; Mishra et al., 2017).
- **Metric-learning approaches** that focus on similarity-based learning to generalize across tasks.
- **Optimization-based methods**, including **Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017)**, which enhance a model's adaptability by optimizing initialization parameters for quick fine-tuning on new tasks.

Domain generalization addresses the issue of **domain shift**, where machine learning models trained on one distribution fail to generalize effectively to different but related distributions (Zhou et al., 2021). Unlike **transfer learning** or **domain adaptation**, domain generalization **does not rely on labeled data from the target domain during training**.

Empirical Risk Minimization (ERM)

ERM (Vapnik, 1999) minimizes the average prediction error across a given dataset, making it a widely used baseline for domain generalization. However, ERM assumes that training and testing data follow the same distribution, which is not always the case in practical applications (Gulrajani & Lopez-Paz, 2020).

Key Domain Generalization Strategies

J. Wang et al. (2021) classify domain generalization methods into:

1. **Data Manipulation:** Techniques such as **data augmentation** (e.g., flipping, rotation, noise addition) and **domain randomization** (altering textures, lighting, or object positions) increase dataset diversity and improve model robustness (Tobin et al., 2017).
2. **Representation Learning:** Learning **domain-invariant representations** minimizes distribution discrepancies between training and testing data, improving generalization to unseen domains (Zhou et al., 2021).
3. **Ensemble Learning & Meta-Learning:** Combining multiple models or adopting **learning-to-learn strategies** enhances a model's ability to generalize across different domains (J. Wang et al., 2021).

Deepfakes

The rapid growth of digital media and the widespread adoption of social networking platforms have led to an unprecedented surge in online images and videos. While these advancements facilitate communication and creativity, they have also given rise to highly sophisticated manipulation techniques powered by deep learning (Malik et al., 2022; Masood et al., 2021). One of the most concerning developments in this field is the emergence of **deepfakes**—AI-generated videos, images, and audio that can convincingly depict individuals saying or doing things they never actually did (Yu et al., 2021).

The term **"deepfake"** originated in 2017, named after a Reddit user who utilized deep learning to superimpose celebrity faces onto videos (Malik et al., 2022). Since then, deepfake technology has expanded beyond just video manipulation to include **both image-based and audio-based forgeries**. These synthetic media creations present a growing challenge, as they blur the distinction between authentic and manipulated content, potentially eroding public trust in digital media (Masood et al., 2021).

Deepfakes have serious implications, ranging from the spread of misinformation (Satariano & Mozur, 2023) to their use in **political manipulation (Meaker, 2023)** and **reputation damage (Mustak et al., 2023)**. As deepfake generation techniques become increasingly advanced, distinguishing between real and manipulated media becomes more difficult (Masood et al., 2021). This growing challenge underscores the urgent need for **effective deepfake detection technologies** to preserve the credibility of digital content.

This chapter explores the creation of **image- and video-based deepfakes**, detailing the techniques used to generate them and the latest detection methods designed to uncover these manipulations.

IV. Proposed Approach and Evaluation

This research integrates **Meta-Learning for Domain Generalization (MLDG)** with **Self-Blended Images (SBIs)** to improve deepfake detection. **The proposed framework is evaluated using the FaceForensics++ dataset (Rössler et al., 2019), employing a leave-one-out cross-validation scheme.**

The experimental evaluation involves:

1. **Training the model on multiple deepfake generation techniques** while excluding one technique for testing.
2. **Comparing the performance of the MLDG + SBI approach against a baseline model using ERM.**
3. **Assessing the model's ability to generalize to unseen deepfake manipulations.**

Meta-learning and domain generalization offer promising solutions for deepfake detection, allowing models to adapt to new manipulations without requiring extensive retraining. By combining **MLDG and SBIs**, this research aims to develop a more **generalizable deepfake detection approach** capable of **detecting previously unseen manipulations**. Future research directions could explore incorporating **temporal information** in video deepfake detection and expanding the approach to **audio-based deepfakes**.

Conclusions

This study examined the challenges of deepfake detection in the context of evolving manipulation techniques. Given the

limitations of traditional detection models, which often struggle to generalize to previously unseen deepfake generation methods, this research explored **Meta-Learning for Domain Generalization (MLDG)** as a potential solution.

By treating each deepfake generation technique as a distinct domain, this study assessed the model's ability to **learn and generalize** to novel manipulation types. A key research question was whether **MLDG could outperform the standard domain generalization baseline, Empirical Risk Minimization (ERM)**. Unlike conventional **meta-learning or few-shot learning** approaches that require some level of adaptation using new data, **MLDG is better suited for real-world scenarios where no prior data from new deepfake methods is available**. This reflects the practical challenges of deepfake detection, where models must remain effective against previously unseen manipulations.

Additionally, this study explored the benefits of **Self-Blended Images (SBIs) for data augmentation**, a technique designed to introduce greater variability into training datasets by generating additional source domains. The underlying hypothesis was that increasing training domain diversity would enable the model to **learn more consistent features** across domains, ultimately improving its ability to detect novel deepfake manipulations.

In baseline experiments, where models were trained without augmented data, **MLDG did not consistently outperform ERM**, despite its theoretical advantages in generalization. A possible explanation is the significant domain shift between training and testing data in deepfake detection, which may limit MLDG's ability to generalize effectively when training data lacks sufficient diversity.

References

[1] Alhogail, A.A., and Alsabih, A. (2021). Machine learning and common dialect preparation are used to distinguish phishing emails. Computer Security, 110, 102414.

[2] Altuncu, E., Franqueira, V.N.L., and Li, S. (2022). Deepfake includes definitions, execution measures, datasets, and a meta-review.

[3] Balaji Y., Sankaranarayanan S., and Chellappa R. (2018). Metareg refers to the use of meta-regularization to achieve spatial generalization. Neural data preparation frameworks.

[4] Batagelj,B.,Peer,P.,Struc,V.,&Dobrisek,S.(2021).How can I correctly recognize face masks for Covid-19 from visual data? Connected sciences.

[5] Bengio, S., Y., Cloutier, J., and Gecsei, J. (1992). Run the show while optimizing synaptic learning. Optimality in Manufactured and Organic Neural Systems, 6–8.

[6] Minister, C.M., and Religious Administrator, H. (2024). Profound learning involves understanding established concepts and ideas. Springer Worldwide Distributing. Reference: https://doi.org/10.1007/978-3-031-45468-4.

[7] Choi, Y.; Choi, M.-J.; Kim, M. S.; Ha, J.-W.; Kim, S.; and Choo, J. (2017).

[8] Stargan developed generative antagonistic systems to understand images across multiple domains. Acknowledgment for the 2018 IEEE/CVF Conference on Computer Vision and Design: 8789-8797. D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi. (2021).

[9] Using efficientnet and vision transformers for video deepfake location. ArXiv: abs/2107.02612. Deepfakes (2020). Deepfakes: https://github.com/deepfakes/faceswap. Deng J., Dong W., Socher R., Li L.-J., Li K., & Fei-Fei L. (2009).

[10] Imagenet is a big, multi-leveled image database. Acknowledgment for the 2009 IEEE Conference on ComputerVisionandDesign,pages248–255. Dong S., Wang J., Ji R., Liang J., Fan H., and Ge Z. (2022). VerifiableCharacterSpillage:

[11] The shaky square is moving forward. Deepfake location generalization. 2023 IEEE/CVF Conference on Computer Vision and Design Acknowledgement (CVPR), 3994–4004. Finn, C. (2022). Space generalization [Part of the course CS]