Literature Review for Ingur Thesis References (Farzane Part)

Reza

September 15, 2024

Contents

Chapter 1

Combining EfficientNet and Vision Transformers for Video Deepfake Detection [1]

1.1 Main Idea

The core idea of the paper is to combine the strengths of EfficientNet, a convolutional neural network (CNN), and Vision Transformers (ViTs) to enhance the ability to detect deepfakes in videos, particularly focusing on faces, which are often the main target of manipulation. The authors argue that CNNs, like EfficientNet, are effective for capturing local features (such as spatial patterns in images), while Vision Transformers are better suited for modeling global information and patterns in larger patches of images. By combining these two architectures, the method can effectively detect subtle visual anomalies introduced by deepfake generation techniques.

1.2 Contributions

1. EfficientNet-ViT Hybrid Models: The paper proposes two mixed convolutional-transformer architectures, namely the Efficient ViT and Convolutional Cross ViT, which combine CNNs (EfficientNet) and ViTs. These models process video frames to detect manipulations in faces and classify them as real or fake.

   (a) Efficient ViT uses EfficientNet B0 as a feature extractor for low-level image features, which are then processed by a Vision Transformer to detect anomalies in face images.

   (b) Convolutional Cross ViT introduces a multi-scale approach by processing both small and large patches using two branches. The outputs from these branches are combined via cross-attention to better capture both local and global artifacts in deepfake videos.

2. Simplified Inference Procedure: A voting-based inference mechanism is introduced to handle multiple faces within a video. Instead of averaging predictions for all faces, the model aggregates scores by actor, and a video is classified as fake if any actor's face is detected as manipulated. This procedure improves robustness in videos where only some actors' faces may have been altered.

3. No Distillation or Ensemble Methods: Unlike state-of-the-art deepfake detection models, the proposed methods do not rely on complex techniques like model distillation (transferring knowledge from a larger model to a smaller one) or ensemble methods (using multiple models to make predictions). This simplifies the training and inference process, while still achieving competitive results.

4. Performance: The models were tested on two widely-used datasets, including the DeepFake Detection Challenge (DFDC) and FaceForensics++. The proposed models achieved high accuracy and F1-scores, approaching state-of-the-art performance. Specifically, the Convolutional Cross ViT with EfficientNet B0 achieved an AUC of 0.951 and an F1 score of 88.0% on the DFDC dataset, close to the best results from more complex models.

5. Public Availability: The authors made the code and model implementations publicly available, allowing others in the research community to reproduce and extend their results.

1.3 Conclusion:

The paper demonstrates that combining EfficientNet with Vision Transformers results in an efficient and effective deepfake detection method. The mixed architecture allows for improved detection by addressing both local and global image features, and the simplified inference mechanism enhances the robustness of the model in real-world scenarios. Despite being simpler than state-of-the-art models, the proposed approach achieves nearly competitive performance without relying on distillation or ensemble techniques.

1.4 Summrize in one paragraph

The paper "Combining EfficientNet and Vision Transformers for Video Deepfake Detection" introduces a novel deepfake detection method that combines EfficientNet, a CNN, with Vision Transformers (ViTs) to capture both local and global image features. The authors propose two architectures: Efficient

ViT and Convolutional Cross ViT, which process video frames of faces to detect manipulations. A voting-based inference mechanism improves the handling of multiple faces in videos. Without relying on complex techniques like distillation or ensemble methods, the models achieve near state-of-the-art performance on datasets like DFDC, with an AUC of 0.951 and F1 score of 88.0%. The approach is efficient and publicly available, making it accessible for further research.

Chapter 2

Towards Solving the DeepFake Problem : An Analysis on Improving DeepFake Detection using Dynamic Face Augmentation [2]

2.1 Main Idea

The main idea of the paper is to improve the performance of deepfake detection models by addressing the problem of overfitting caused by oversampled datasets. To achieve this, the authors propose a new data augmentation technique called Face-Cutout, which dynamically removes parts of the face during training based on facial landmarks. This method helps the model focus on relevant regions of the face where manipulation is more likely to occur, improving its ability to generalize and detect deepfakes across different datasets. The paper also introduces guidelines for face clustering and dataset preprocessing to prevent data leakage and ensure that models are evaluated more accurately.

2.2 Contributions

1. Identification of Dataset Issues: The authors perform a detailed analysis of popular deepfake datasets, including DFDC, FaceForensics++, and Celeb-DF, highlighting that these datasets are oversampled and suffer from a lack

of variation in real faces. They show that this leads to overfitting, where models memorize faces rather than learning to detect deepfake features.

2. Face-Cutout Augmentation Technique: The core contribution of the paper is the introduction of Face-Cutout, a novel data augmentation method that selectively cuts out regions of the face based on facial landmarks. This technique focuses on areas where deepfake manipulations are more likely to occur, helping to improve the model's ability to learn and detect manipulated features rather than irrelevant areas of the face. This method significantly reduces overfitting and improves model generalization.

3. Improved Training Preprocessing: The authors propose a preprocessing guideline that involves face clustering before training deepfake detection models. Instead of randomly splitting datasets into training and test sets, they suggest splitting the data based on unique faces to avoid data leakage. This ensures that the models do not train and test on the same faces, improving the robustness of evaluation results.

4. Quantitative Improvements: The paper presents experimental results showing that models trained with Face-Cutout achieve significant improvements in performance across multiple datasets (DFDC, FaceForensics++, and Celeb-DF). Face-Cutout reduces LogLoss by 15.2% to 35.3% compared to baseline models and other occlusion-based augmentation methods like Random-Erasing. The method enhances both EfficientNet-B4 and Xception models, increasing their ability to generalize to unseen data.

5. Interpretability of Models: To validate their claims of reduced overfitting, the authors use Grad-CAM to visualize how the models focus on different regions of the face. The results show that models trained with Face-Cutout successfully focus on manipulated areas of the face, while baseline models often highlight irrelevant regions, indicating overfitting.

6. Generalization Across Datasets: The authors demonstrate that Face-Cutout is effective across different datasets, improving performance on both individual and combined datasets. They also show that it can be integrated into existing deepfake detection pipelines without requiring significant changes to the models.

2.3 Conclusion

The paper contributes to solving the deepfake detection problem by identifying the root causes of overfitting in current models and proposing an effective solution through Face-Cutout augmentation. The method improves the generalizability of deepfake detection models, reduces overfitting, and enhances their ability to detect manipulated faces across different datasets. The authors also provide practical preprocessing guidelines to prevent data leakage and ensure more robust evaluation. The proposed method is shown to be effective across various architectures and datasets, making it a valuable addition to deepfake detection techniques.

2.4 Summrize in one paragraph

The paper ”Towards Solving the DeepFake Problem: An Analysis on Improving DeepFake Detection Using Dynamic Face Augmentation” addresses the issue of overfitting in deepfake detection models caused by oversampled datasets with limited facial variation. To solve this, the authors propose a new data augmentation method called Face-Cutout, which selectively removes regions of the face based on facial landmarks to help models focus on areas likely to be manipulated. They also introduce a preprocessing guideline using face clustering to prevent data leakage during training. Experimental results show that Face-Cutout significantly improves model performance, reducing overfitting and increasing generalizability across multiple datasets. The method can be easily integrated into existing deepfake detection pipelines, leading to more robust and accurate detection.

Chapter 3

Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis [3]

3.1 Main Idea

The main idea is to assess the potential of self-supervised Vision Transformers (ViTs) in detecting deepfakes, comparing them against both supervised ViTs and ConvNets. The authors examine two approaches: using frozen ViT backbones as feature extractors and partially fine-tuning the final transformer blocks. They show that SSL pre-trained ViTs, particularly DINO and MAE models, achieve superior performance in deepfake detection compared to conventional ConvNets and supervised ViTs. The use of SSL on ViTs offers improved generalizability, robustness, and explainability of the detection process, making them more effective with smaller training datasets.

3.2 Contributions

1. Comparative Analysis: The paper provides a thorough comparison of SSL pre-trained ViTs, supervised ViTs, and ConvNets for deepfake detection. It highlights the advantages of SSL pre-training, especially in improving performance and explainability when working with limited data.

2. Two Approaches for ViTs: The authors propose and evaluate two methods for using ViTs in deepfake detection: (1) using frozen ViT backbones as multi-level feature extractors with simple classifiers, and (2) partially fine-tuning the last transformer blocks for better adaptation to the task.

3. Improved Generalization and Explainability: Through fine-tuning, SSL pre-trained ViTs, particularly DINOv2 and MAE, demonstrate better generalization across various deepfake datasets. The attention mechanism of transformers also enables better explainability by identifying relevant regions of the face, such as the eyes, nose, and mouth, where deepfake artifacts are often found.

4. Experimental Validation: The paper validates its findings through extensive experiments on multiple deepfake datasets, showing that SSL pre-trained ViTs outperform ConvNets and supervised ViTs in both seen and unseen test sets. They also perform well in cross-dataset evaluations, demonstrating robustness to different deepfake generation techniques.

5. Insight into Self-Supervised Learning: The study shows that SSL methods like DINO and MAE produce strong feature representations for deepfake detection, even when pre-trained on datasets unrelated to deepfakes, underscoring the importance of SSL in vision tasks.

3.3 Conclusion

The paper concludes that SSL pre-trained ViTs are highly effective for deepfake detection, outperforming traditional ConvNets and supervised ViTs in generalizability and explainability. By fine-tuning the last few transformer blocks, these models adapt well to deepfake detection tasks and provide more interpretable results, making them a valuable tool for digital forensics and media security.

3.4 Summrize in one paragraph

The paper "Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis" investigates the effectiveness of self-supervised Vision Transformers (ViTs) compared to supervised ViTs and ConvNets for detecting deepfakes. It shows that self-supervised pre-trained ViTs, particularly models like DINO and MAE, offer superior generalization and explainability for deepfake detection, especially when fine-tuned on small datasets. The study evaluates two approaches: using frozen ViTs as feature extractors and partially fine-tuning their final layers. Experimental results demonstrate that self-supervised ViTs outperform ConvNets and supervised ViTs, providing better performance and more interpretable results by focusing on relevant facial regions where deepfake artifacts typically occur.

Chapter 4

Xception: Deep Learning with Depthwise Separable Convolutions [4]

4.1 Main Idea

The main idea of the paper is to enhance the efficiency of deep learning models by introducing the Xception architecture, which replaces the complex Inception modules with depthwise separable convolutions. This architectural change separates spatial and cross-channel correlations, leading to a more efficient use of model parameters while maintaining or improving performance on large-scale image classification tasks. The architecture also includes residual connections, further aiding in performance and convergence. The authors demonstrate that Xception slightly outperforms Inception V3 on the ImageNet dataset and significantly outperforms it on a much larger dataset (JFT) without increasing model size.

4.2 Contribution

1. Introduction of Depthwise Separable Convolutions: Xception introduces a novel approach by replacing the traditional convolutions and Inception modules with depthwise separable convolutions, improving computational efficiency. Depthwise separable convolutions consist of a depthwise convolution (which applies a spatial convolution independently over each input channel) followed by a pointwise convolution (a 1x1 convolution), decoupling spatial and cross-channel correlations.

2. Simplified Architecture with Residual Connections: Xception is structured as a linear stack of depthwise separable convolution layers with residual

connections, making it easier to design and modify compared to the more complex Inception models. The use of residual connections accelerates convergence and improves the final classification performance.

3. Performance Gains Without Increasing Parameters: The Xception architecture achieves better performance on image classification tasks compared to Inception V3, while maintaining a similar number of parameters. On the ImageNet dataset, Xception shows a marginal improvement over Inception V3, and on the much larger JFT dataset, it achieves a 4.3% relative improvement in mean average precision for top 100 predictions (MAP@100). This demonstrates that the performance gains are due to a more efficient use of parameters rather than an increase in model capacity.

4. Analysis of Non-Linearity in Depthwise Separable Convolutions: The authors explore whether including non-linear activation functions (ReLU or ELU) between the depthwise and pointwise convolutions improves performance. They find that omitting these non-linearities leads to faster convergence and better final performance, in contrast to prior findings with Inception modules.

5. Generalization to Larger Datasets: Xception's effectiveness is validated on the large-scale JFT dataset, showing that it generalizes well to larger, more complex datasets. The architecture's performance on JFT demonstrates its ability to handle large-scale image classification tasks with thousands of classes and millions of images, achieving significant performance improvements over Inception V3.

4.3 Conclusion

The paper concludes that Xception represents a significant improvement over the Inception architecture, providing a more efficient model with better performance in image classification tasks, without increasing the number of parameters. By using depthwise separable convolutions and residual connections, Xception demonstrates the advantages of decoupling spatial and cross-channel correlations in convolutional neural networks. The authors suggest that this approach could become a cornerstone of future neural network architecture designs, as it balances simplicity, efficiency, and performance.

4.4 Summrize in one paragraph

The paper ”Xception: Deep Learning with Depthwise Separable Convolutions” introduces the Xception architecture, an improvement over the Inception model that replaces its complex modules with depthwise separable convolutions. This architecture decouples spatial and cross-channel correlations in convolutional neural networks, leading to more efficient use of parameters without increasing model size. Xception integrates residual connections, enabling faster convergence and better performance. The model outperforms Inception V3 on both the ImageNet and JFT datasets, particularly showing significant gains in large-scale image classification tasks. The study concludes that Xception's simpler yet more effective design could serve as a foundation for future deep learning models.

Chapter 5

Very Deep Convolutional Networks for Large-Scale Image Recognition [5]

5.1 Main Idea

The main idea of the paper is to explore how increasing the depth of ConvNets while using small 3x3 convolution filters can improve image classification performance. By stacking more convolution layers with smaller filters, the model can capture more complex patterns in the data, while maintaining a reasonable number of parameters. This deeper architecture, with layers ranging from 16 to 19, proved to be highly effective for large-scale image recognition tasks, demonstrating that depth plays a critical role in the success of ConvNets.

5.2 Contributions

1. Deeper ConvNet Architecture (VGGNet): The paper introduces a family of deeper ConvNet architectures with 16 and 19 layers, which outperform shallower networks used in previous models. These architectures use small 3x3 convolution filters throughout the network to capture fine details in images while keeping the model's complexity manageable. This deep architecture was critical in achieving state-of-the-art performance in image recognition.

2. Use of Small 3x3 Convolutions: Instead of using larger convolution filters (such as 7x7 or 11x11, as in previous models), the authors use multiple 3x3 convolution layers stacked together. This design increases the non-linearity of the model while allowing it to effectively capture both local and global image features, leading to improved accuracy without a dramatic

increase in parameters.

3. Residual Insights on Architecture Design: The paper systematically increases network depth and demonstrates that deeper models consistently outperform shallower ones, provided that they are appropriately regularized and trained on large datasets. The results suggest that depth is a crucial factor in ConvNet performance, leading to better feature representation and improved classification accuracy.

4. State-of-the-Art Results on ImageNet: VGGNet achieved state-of-the-art results in the ImageNet 2014 challenge, securing second place in the classification task with a top-5 error rate of 7.3% and first place in the localization task. The model's ability to generalize well to other datasets, such as PASCAL VOC and Caltech, further validated its efficacy.

5. Public Availability of Models: To promote further research, the authors made their best-performing ConvNet models publicly available. This allowed the broader computer vision community to build on the success of VGGNet, facilitating its adoption and adaptation for various tasks.

6. Feature Transfer to Other Tasks: The paper also demonstrates that the deep features learned by the VGGNet models transfer well to other image classification tasks. Even without fine-tuning, these features perform exceptionally well on datasets like PASCAL VOC and Caltech-101/256, further proving the versatility of deep representations.

5.3 Conclusion

The paper concludes that increasing the depth of convolutional networks with small convolution filters leads to substantial improvements in image classification accuracy. The proposed VGGNet architecture set new benchmarks in large-scale image recognition, outperforming previous state-of-the-art models. The success of these deep networks underscored the importance of depth in deep learning and established VGGNet as a foundational architecture in the field of computer vision.

5.4 Summrize in one paragraph

The paper ”Very Deep Convolutional Networks for Large-Scale Image Recognition” by Karen Simonyan and Andrew Zisserman introduces the VGGNet architecture, which significantly improves image classification performance by increasing the depth of convolutional neural networks (up to 19 layers) using small 3x3 convolution filters throughout the network. This deeper architecture, designed to capture complex patterns while maintaining a manageable number of parameters, outperformed previous models, securing top positions in the ImageNet 2014 competition. The paper demonstrates that depth is crucial for improving image classification accuracy and that the features learned by these deep networks transfer well to other tasks like object detection and localization. The success of VGGNet has made it a widely adopted architecture in computer vision research.

Bibliography

[1] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," *CoRR*, vol. abs/2107.02612, 2021. [Online]. Available: https://arxiv.org/abs/2107.02612

[2] S. Das, S. Seferbekov, A. Datta, M. S. Islam, and M. R. Amin, "Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3776–3785.

[3] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Exploring self-supervised vision transformers for deepfake detection: A comparative analysis," 2024. [Online]. Available: https://arxiv.org/abs/2405.00355

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016. [Online]. Available: http://arxiv.org/abs/1610.02357

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556