Ingure Thesis summrise (reza)

Reza Adinepour
School of Electrical and
Computer Engineering
Amirkabir University of Technology
Tehran, Iran 30332–0250
Email: adinepour@aut.ac.ir

Farzane Arzaghi
School of Electrical and
Computer Engineering
Amirkabir University of Technology
Tehran, Iran 30332–0250
Email:farzaghi@aut.ac.ir

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678–2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

***Abstract*—The abstract goes here.**

I. Introduction

A. Advances in AI and Deepfakes

The development of Artificial Intelligence (AI) and deep learning technologies has significantly advanced computer vision and image generation capabilities. A key outcome of these advancements is the enhancement of deepfakes, which are AI-generated images and videos that can swap faces with a high degree of realism. While this technology has brought value to creative fields such as filmmaking and special effects, it has also raised ethical concerns. The ability to manipulate videos to make people appear to say or do things they never did presents serious risks to privacy, security, and trust in digital media. Therefore, detecting and mitigating deepfakes has become essential to protect potential victims of such manipulations.

B. Early Approaches in Deepfake Detection

Initially, Convolutional Neural Networks (CNNs) were used to identify deepfakes, as they had proven successful in various computer vision tasks. Early CNN-based models showed effectiveness in identifying first-generation deepfakes. However, as datasets became larger and more complex, these models required further advancements to improve both their accuracy and capacity to handle diverse and sophisticated deepfakes.

C. Shifting to Transformers in Vision Research

In 2020, researchers began exploring the potential of transformer-based models for computer vision tasks, inspired by their success in Natural Language Processing (NLP). Transformers, particularly those with the self-attention mechanism, became a preferred model due to their ability to capture long-range dependencies within data. The Vision Transformer (ViT) was introduced as an alternative architecture that, unlike CNNs, processes images as a series of patches rather than pixel-based grids. This novel approach leveraged large datasets and pre-training, allowing ViTs to perform on par or even better than state-of-the-art CNN models in some cases.

D. Challenges with Vision Transformers in Deepfake Detection

Despite their success, Vision Transformer-based models still face certain challenges in deepfake detection, particularly in terms of resource requirements. ViTs often demand high computational power and memory, which becomes problematic in real-time or resource-constrained environments, such as on edge devices. To address these bottlenecks, researchers continue to explore methods that balance accuracy with efficiency.

E. Model Compression Techniques

Recent efforts have focused on reducing the computational cost of models through techniques like quantization, knowledge distillation, and pruning. These approaches compress model sizes by lowering the precision of weights without significantly altering the architecture. Among these, post-training quantization (PTQ) has been widely adopted due to its simplicity and its ability to minimize memory use while speeding up model inference. However, these methods often lead to some loss in accuracy, especially at lower precision levels.

F. Quantization-Aware Training (QAT)

To mitigate the loss of accuracy, Quantization-Aware Training (QAT) has been introduced. This approach allows models to adapt to reduced precision during the training phase itself, leading to better performance during inference, albeit with higher resource demands during training.

G. BitLinear Layer and Efficient Deepfake Detection

More recently, BitNet introduced the BitLinear layer, which replaces traditional linear layers with ternary-weighted layers (-1, 0, +1). This innovation reduces both memory and energy consumption, making transformers more scalable and efficient. Motivated by the need for more resource-efficient models, this thesis investigates how BitLinear layers can enhance Vision Transformers for deepfake detection, especially in resource-constrained settings.

H. Key Research Questions

This work seeks to address the following questions:

1) How can Vision Transformers be adapted to effectively detect deepfakes in environments with limited resources?

2) What impact do BitLinear layers have on the performance, size, and speed of Vision Transformers?
3) How do quantization-aware models compare to post-training quantized models in deepfake detection tasks?

I. Contributions of This Thesis

To answer these questions, this thesis aims to:

1) Implement the BitLinear layer in custom networks, building on the BitNet framework.
2) Pre-train two Vision Transformer models, one using full-precision linear layers and another utilizing BitLinear layers, on ImageNet-1k.
3) Fine-tune these models on multiple deepfake datasets (e.g., FF++, Celeb-DF), and compare their detection performance.
4) Quantize a baseline model to evaluate the effectiveness of QAT and PTQ techniques in the context of deepfake detection.

II. Literature review

A. Dataset

Modern deepfake generation techniques have evolved from early methods like 3D morphable models and autoencoders to more advanced techniques, including GANs (e.g., StyleGAN, FSGAN) and diffusion models, which have significantly improved the realism of synthetic faces. To support the advancement of deepfake detection, various datasets have been developed over time, categorized into three generations. The first generation includes small datasets like UADFV, DF-TIMIT, and FaceForensics++, which provided early benchmarks. The second generation, featuring Google DFD, DFDC Preview, and Celeb-DF, introduced higher-quality deepfakes, with Celeb-DF being one of the most widely used. The third generation, including the DeepFake Detection Challenge (DFDC) and Deeper Forensics, expanded the dataset size and diversity, addressing limitations such as the number of swapped identities.

B. CNNs

Deepfake detection is fundamentally an image classification task, requiring models to distinguish between real and manipulated faces. Early research primarily used Convolutional Neural Networks (CNNs) due to their success in image classification, with MesoNet being one of the first deep learning methods developed for this purpose in 2018. MesoNet's compact architecture targets mid-level features for efficient detection, while MesoInception extended its capabilities by capturing multiscale features.

Li and Lyu advanced detection by using larger CNN models (e.g., ResNet) to identify warping artifacts in deepfake videos. They trained their models on UADFV and DF-TIMIT datasets, showing that ResNet50 performed best. Another notable approach involved combining VGGFace with ResNet50, achieving high accuracy on datasets like FF++ and Celeb-DF.

Recent work has also compared models using full-face inputs versus specific facial regions. Results consistently show that models using entire face inputs outperform region-based models. Finally, the DeepFake Detection Challenge revealed that ensemble learning and frame-by-frame classification generally produce better results. Top-performing solutions utilized models like EfficientNet and XceptionNet, with techniques such as augmentations and MixUp for improved accuracy.

C. Transformer

III. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.