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

I. Introduction

In recent years, the rise of AI-generated deepfakes has posed significant challenges to the integrity of digital media, making it increasingly important to develop reliable detection methods. Traditional approaches, such as Convolutional Neural Networks (CNNs), have been instrumental in detecting manipulated media by focusing on local features within images. However, as deepfake technology evolves, these methods are proving inadequate for capturing the more sophisticated manipulations that are now possible. Recognizing this limitation, I turned my attention to Vision Transformers (ViTs), which offer a more advanced capability by analyzing images as sequences of patches, thereby allowing for both local and global feature extraction. The core objective of this research is to explore how Vision Transformers can be effectively utilized for deepfake detection while addressing their inherent computational demands. To achieve this, I propose a novel integration of ViTs with BitLinear layers, a quantization technique that significantly reduces the model's computational complexity without sacrificing accuracy. This approach is particularly relevant for real-time applications and resource-constrained environments, where both performance and efficiency are critical.

II. literature review

A. CNNs

Convolutional Neural Networks (CNNs) have long been the foundation of many deepfake detection models due to their success in image classification tasks. However, as deepfake technology advances, the limitations of CNN-based models become more pronounced, particularly in their ability to handle the increasingly sophisticated manipulations present in second-generation deepfakes.

One of the major drawbacks of CNNs is their reliance on local feature extraction, which, while effective for detecting low-level artifacts in earlier deepfakes, often fails when dealing with high-quality forgeries. The paper by Li and Lyu [17], for instance, highlights how traditional CNN models, even when enhanced with advanced architectures like ResNet, struggle to detect deepfakes when the manipulations involve subtle, high-resolution adjustments that do not produce obvious artifacts.

Moreover, the study by Korshunov and Marcel [22] demonstrates that CNN-based systems, such as those relying on VGG or Facenet architectures, are particularly vulnerable to high-quality deepfakes. The research shows that these systems were easily deceived by deepfakes in 85-95% of cases, underscoring their inadequacy in maintaining high accuracy when faced with realistic forgeries.

The inherent architecture of CNNs also limits their capacity to capture global context within an image, which is increasingly important as deepfakes become more sophisticated. For example, CNNs typically process images by focusing on small, localized regions, which makes them less effective at recognizing broader inconsistencies that span across an entire image or video frame. This limitation is evident when comparing CNNs with Vision Transformers (ViTs), which excel at capturing both local and global features by treating an image as a sequence of patches. The combination of CNNs with ViTs in models like the Convolutional Vision Transformer (CViT) shows significant improvement in detection accuracy, indicating that CNNs alone are insufficient for handling the complexity of modern deepfakes [35].

Another critical issue with CNN-based models is their generalization capability. As shown in the literature, CNNs often require extensive fine-tuning when applied to different datasets, which reduces their practicality in real-world applications where deepfake characteristics can vary widely. The reliance on specific dataset characteristics makes CNNs less adaptable to new or unseen deepfake variations, as demonstrated in several studies where CNNs performed well on training data but failed to generalize to more challenging, unseen examples [40].

In conclusion, while CNNs have been instrumental in the early stages of deepfake detection, their limitations in handling advanced manipulations, capturing global image context, and generalizing across different datasets highlight the need for more robust approaches. The integration of Vision Transformers and quantization techniques, as proposed in this thesis, addresses many of these shortcomings by providing a more comprehensive and efficient solution for deepfake detection.

B. Vision Transformers in Deepfake Detection

Vision Transformers (ViTs) have been pivotal in advancing deepfake detection. ViTs process images by dividing them into patches, treating each patch as a token in a sequence, similar to words in a sentence in natural language processing tasks. This approach allows ViTs to capture complex relationships across different parts of an image, making them particularly effective for detecting subtle manipulations typical in deepfakes.

The thesis proposes the LinViT and BitViT architectures, which leverage ViTs for feature extraction combined with quantization techniques to optimize performance. This is comparable to the work by Coccomini et al. [2], who used a hybrid approach by combining EfficientNet with Vision Transformers. In their model, EfficientNet handles local feature extraction while the ViT captures global features, similar to how the thesis integrates ViTs with quantization for efficiency.

Another study by Heo et al. [3] explored the use of a Data-Efficient Image Transformer (DeiT) as the primary feature extractor for deepfake detection. This model was optimized through a student-teacher distillation process, allowing the ViT to learn more efficiently from a larger CNN-based model. The thesis's approach of using BitLinear layers can be seen as an alternative efficiency improvement, focusing on reducing the computational overhead during both training and inference, while Heo's model emphasizes efficient training through distillation.

C. Quantization Techniques

Quantization is critical for deploying large models like Vision Transformers in real-world applications, especially in environments with limited computational resources. Quantization reduces the bit-width of model weights, thus lowering memory usage and speeding up inference. However, this often comes at the cost of reduced accuracy, particularly in models requiring fine-grained analysis, such as those used for deepfake detection.

This paper employs BitLinear layers, which are a form of quantization that uses ternary weights −1, 0, +1, allowing for substantial reductions in model size and computational requirements without significant loss in accuracy. This approach is particularly novel compared to standard Post-Training Quantization (PTQ) methods, which map high-precision weights to lower bit-widths after training. As discussed by Wang et al. [76] in their work on BitNet, this method allows the model to maintain performance comparable to full-precision models while being much more efficient.

The literature also discusses Quantization-Aware Training (QAT), a technique that quantizes the model during training to allow it to adapt to lower precision weights. While QAT generally results in better accuracy retention compared to PTQ, it is also more resource-intensive. The paper's use of BitLinear layers offers a middle ground, providing many of the benefits of QAT without the associated training overhead. This is evident when comparing the performance of the BitViT model proposed in the paper with models from the literature that use QAT, such as the work by Dettmers et al. [22], where the trade-offs between resource usage and accuracy are clear.

The proposed BitViT model in the thesis is evaluated on standard deepfake detection benchmarks, such as the Celeb-DF dataset. The thesis reports competitive accuracy rates, which align well with those achieved by other state-of-the-art models that also utilize Vision Transformers.

Abstract—The abstract goes here.

References

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

[2] D. A Coccomini, N. Messina, C. Gennaro, and F. Falchi. Combining EfficientNet and Vision Transformers for Video Deepfake Detection

[3] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. Deepfake de- tection scheme based on vision transformer and distillation

I have not written all the references