MSc Artificial Intelligence
Master Thesis

Your Title Here

by
Ingur Veken
11886366

August 17, 2024

48 EC
Period in which the research was carried out

Supervisor:
Dr A Person

Examiner:
Dr A Person

Second reader:
Dr A Person

Universiteit van Amsterdam

Contents

Abstract

1 Introduction

Recent advances in Artificial Intelligence (AI) and deep learning have led to considerable progress in the field of computer vision and image synthesis. In particular, these advancements have greatly improved the quality of deepfakes - AI techniques that enable realistic-looking face swapping in images and videos. However, while this technology has shown promise in various creative applications, such as in movie localization and special effects, it has also raised significant concerns. The widespread ability to manipulate videos, making individuals appear to say or do things they never said or did, poses severe threats to privacy, security, and trust in digital media [13]. For this reason, deepfake detection methods have become critical to protect those susceptible to this technology.

Early attempts to detect deepfakes predominantly used Convolutional Neural Networks (CNNs), due to their successful application in various computer vision tasks. Various CNN architectures showed promising results on first generation deepfake datasets [1, 14, 30, 63]. However, as the quality, quantity and diversity of these datasets increased, more powerful models need to be developed.

In 2020, computer vision research began shifting towards transformer-based models following its success in large language models (LLMs) for Natural Language Processing (NLP). Transformers quickly became the de facto standard architecture in NLP, as the self-attention mechanism allowed for capturing long-range dependencies and complex contexts [69]. The scaling successes of transformers led to the development of the Vision Transformer architecture, which uses raw image patches rather than text tokens as input [24]. To address the lack of inductive biases found in CNN architectures, these models largely relied on larger datasets for pre-training. This allowed Vision Transformers to achieve performance comparable to state-of-the-art CNN models, sometimes even outperforming them [67].

In deepfake detection, ViT-based architectures either use transformers after a CNN to further refine the extracted features, or replace the CNN entirely. Nonetheless, the adoption of Vision Transformers as feature extractors for deepfake detection still faces various challenges. As ViTs continue to improve and grow in size, the increasing computational and memory requirements create a significant bottleneck. This is especially noticeable in resource-constrained environments, such as for edge devices or real-time applications [25].

Recently, different approaches to model compression have been researched, including quantization, knowledge distillation and pruning [50, 66, 78]. Quantization maps high precision model weights to lower bit-widths without changing the model architecture, and is regarded as one of the most effective approaches to model compression [27]. Most quantization methods are done post-training (PTQ), as this does not require any changes to existing training pipelines, reduces the memory footprint and can speed up model inference drastically [49]. Unfortunately, these methods also result in accuracy loss, especially at lower precisions [20, 70].

Quantization-Aware Training (QAT) aims to counter this loss in accuracy by emulating quanti-

zation during training [70]. This way the model learns to adapt to the reduced precision, which allows it to maintain higher accuracy at inference-time at the cost of being more time-cosuming and resource intensive to train [25].

More recently, BitNet introduced the BitLinear layer, a custom linear layer that uses ternary $\{-1, 0, +1\}$ weights, designed for scalable and stable Transformer LLMs. BitNet replaces all linear layers with BitLinear layers that are trained using QAT, to significantly reduce memory and energy consumption compared to their higher-precision counterparts [51, 70].

Motivated by the need for more accessible and efficient detection models, the goal of this thesis is to investigate the application of BitLinear layers to Vision Transformers for deepfake detection. Specifically, this thesis attempts to answer the following research questions:

- How can Vision Transformers be efficiently adapted for deepfake detection in resource-constrained environments?
- What is the impact of using BitLinear layers in Vision Transformers on the model size, speed and detection performance?
- How do quantization-aware trained models compare to post-training quantized models in the context of deepfake detection?

To address these questions, we aim to make the following contributions TODO: rewrite these:

- We implement the BitLinear layer for training custom networks, based on the BitNet paper.
- We pre-train two Vision Transformers, a baseline using full-precision linear layers and our implementation of BitLinear layers, on ImageNet-1k.
- We fine-tune these models on multiple deepfake datasets, FF++, Celeb-DF (and maybe a DFDC subset), and compare the results.
- We quantize the baseline model to compare the effectiveness of the QAT and PTQ approach.

The remainder of this thesis will be structured as follows: TODO: briefly explain structure

2 | Literature review

In this chapter, we discuss past research that focuses on Deepfake detection methods, CNN- and transformer-based architectures and challenges. We highlight differences and similarities among models, and the problems they address.

2.1 Datasets

TODO: fix this introduction below: Modern deepfake generation methods use different deep learning techniques to manipulate or synthesize faces in photos or videos. Early face-swapping techniques mainly relied on 3D morphable face models (3DMM) or autoencoder based methods. Later, Generative Adversarial Networks (GANs) such as StyleGAN [40] and FSGAN [54] allowed for realistic synthetic face generation and face swapping respectively. More recently, diffusion models (?) have also shown promising results in image synthesis tasks, allowing for even more realistic fake faces to be generated.

Multiple datasets have been released over the past few years to support the advancement of deepfake detection models. These datasets can be categorized into three generations based on their release date, size and synthesis quality. The first generation consists of small datasets such as UADFV [73] and DF-TIMIT [42], as well as FaceForensics++ (FF++) [56], which provided the initial benchmarks for early deepfake detection models. Of these datasets, FF++ is the largest and most widely explored dataset, containing 1000 real and 4000 counterfeit videos. The second generation dataset includes Google DFD [26], DFCD-Preview [22] and Celeb-DF [48], and greatly improved the visual quality of the deepfake videos, providing more challenging data for detection models. Celeb-DF is the most used dataset of this generation, offering 590 real and over 5000 fake videos, corresponding to more than 2 million unique frames. Finally, the third generation consists of the DeepFake Detection Challenge (DFDC) [23] and Deeper Forensics [39] datasets, increasing the amount of data and variety even further. Most importantly, these datasets address the most common problem with the first two generations, namely the limited amount of swapped identities [23]. For example, Celeb-DF uses videos from 59 different celebrities, whereas DFDC contains 100,000 videos of 3426 paid actors, which amounts to tens of millions of frames.

2.2 Convolutional Networks

Fundamentally, deepfake detection is an image classification problem, where models need to be able to detect slight differences between fake and real faces. For this reason, initial deep learning research in Deepfake detection used Convolutional Neural Networks (CNNs) to identify manipulated images/videos, due to their widespread success in image classification tasks [7].

In 2018, MesoNet [1] presented one of the first deep learning approaches to deepfake detection.

MesoNet uses a very compact CNN architecture designed for efficiency, and focuses on detecting mesoscopic (mid-level) features in images. The authors also introduced MesoInception, which uses inception modules to capture multiscale features to further increase its detection methods [62].

In contrast, Li and Lyu [47] used large conventional CNN models to detect specific warping artifacts observed in early deepfake videos. This work took advantage of the computational limitations of early deepfake algorithms, which only synthesized face images of fixed sizes, and applied affine warping to match the target faces. The warping process results in visible artifacts, as a result of the resolution inconsistency between the warped face area and surrounding context. The authors experimented with four different CNN models based on the VGG16, ResNet50, ResNet101 and ResNet152 architectures [30, 58]. The models were trained using a self-generated dataset, whereas evaluation was done on the UADFV and DF-TIMIT datasets. Furthermore, they also trained and compared their results to the MesoNet and MesoInception models on the same datasets. Their results showed that the ResNet50 model outperformed all other models on both datasets.

Wang et al. [71] combined the VGGFace face detection model [8] with ResNet50 as backend architecture for their model. Their approach monitors the internal activations of the neurons at various layers and captures statistical data, which is then used to train the classifier. This model achieved AUC scores of 98.5% on FF++, 66.8% on Celeb-DF and 68% on the DFDC dataset.

Another study compared selecting the entire face as input to detection models versus only selecting specific regions [65]. Here, the XceptionNet [14] architecture was used to create the detection models, and the results were evaluated on datasets from both the first (UADFV, FF++) and second generation (Celeb-DF, DFDC-Preview). Evaluation showed significant realism improvements for specific facial regions between the first and second generation datasets, especially the nose, mouth and edge of face regions. The results also show that full face input models outperform all specific facial region based models across all datasets.

For the DeepFake Detection Challenge [23], the authors performed a meta analysis of the top submissions to the detection challenge. The top-5 winning solutions all used ensemble learning, which combines multiple model outputs to further increase the accuracy. Moreover, frame-by-frame classification models also generally performed better when compared to multi-frame models. The first-place submission used MTCNN [77] for face detection and an ensemble of seven models based on the EfficientNet B7 [63] architecture for feature encoding. Additionally, the training made heavy use of different augmentations, and also included dropping structured parts of faces to further improve generalizability [19]. The second solution was based on the XceptionNet and EfficientNet B3 architecture, and a WS-DAN model [35] for augmentation. The third place submission also used an ensemble of EfficientNet models and made use of the MixUp augmentation [76].

2.3 Vision Transformers

The modern Transformer architecture was introduced in 2017 by Vaswani et al. [69] and primarily designed for Natural Language Processing (NLP) tasks. The self-attention mechanism allows for capturing long-range dependencies and understanding complex contexts. However, while Transformers quickly became the de facto standard for NLP tasks, this design initially saw limited application in Computer Vision tasks. This is because naive application of self-attention on high-dimensional data such as images would be too computationally and memory

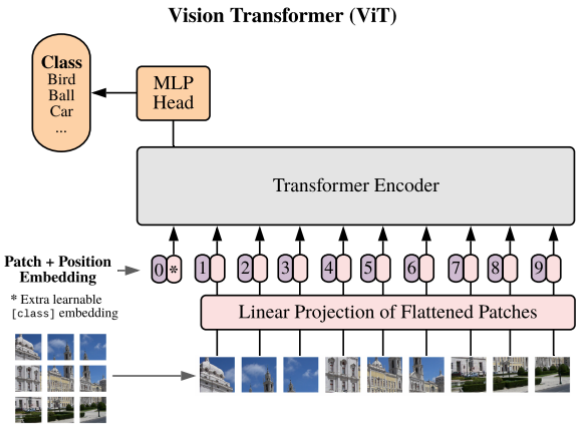

Figure 2.1: Vision Transformer (ViT) architecture. Images are split into fixed-sized patches, which are linearly projected. Positional encoding is applied to the resulting sequence of projected patches, which are then processed by a Transformer encoder. Figure from [24].

intensive, as the cost of self-attention scales quadratically ($O(n^2)$) with the dimensionality of the data. Additionally, Transformers lack some of the inductive biases found in CNN architectures, including locality and translation equivariance [24].

Despite these initial limitations, Dosovitskiy et al. [24] introduced the Vision Transformer (ViT) in 2020 after the scaling successes of Transformers in NLP. ViT models work by splitting images into linearly embedded fixed-sized patches, and consider each patch a token in a sequence (see Fig. 2.1). This patch-based approach allows for more efficient application of the self-attention mechanism, whilst still capturing complex relationships between different parts of the image. To address the lack of inductive biases found in CNN architectures, ViT models rely on pre-training with large datasets. This allows the architecture to achieve performance comparable to state-of-the-art CNNs on multiple image recognition benchmarks.

In the context of deepfake detection, Vision Transformers are mainly used in two ways: Following a CNN or as a replacement for CNNs as the main feature extractor [53]. Wodajo and Atnafu [72] introduced the CViT model which uses a CNN model based on the VGG architecture [58] as feature learning component. This is motivated by the fact that CNNs excel at learning local features, while Transformers can learn from local and global feature maps. The combined capacity enables their model to better correlate relationships between non-local features.

Later, Coccomini et al. [15] expanded upon this architecture by replacing the VGG CNN with an EfficientNet B0 as the local feature extractor. The authors also replace the ViT with a CrossViT based multiscale architecture [11]. CrossViT uses a dual-branch transformer approach that processes both small and large patch tokens in two separate branches and uses cross-attention to exchange information between the different scales. This allows the model to better detect artifacts introduced by deepfakes both on a local and global level.

In contrast, Heo et al. [33] used a Data-Efficient Image Transformer (DeiT) [66] as the main feature extractor. DeiT uses a student-teacher strategy and introduces a special distillation token. In this case, an EfficientNet B7 is used to produce soft labels, which the ViT student model learns to replicate. This method provides an extra learning signal to the student model, which speeds up the training and further increases the accuracy of the model, allowing it to be trained using less data.

Recent work also investigated the efficacy of self-supervised learning (SSL) on pre-trained vision

transformers for deepfake detection [53]. Self-supervised learning is a method for learning robust representations from unlabeled data and played an important role in the success of Transformers in NLP. In language modelling, self-supervised training objectives provide a more complex learning signal than supervised objectives of a single label per sentence [55]. Similarly, Caron et al. [9] argue that image-level supervision often reduces the rich visual information contained in an image to a single concept selected from a predefined category set. The authors applied SSL to DeiT, which led to the development of DINO, extending knowledge distillation to work without explicit labels. Nguyen et al. [53] show that fine-tuning DINO for deepfake detection can lead to superior performance compared to fine-tuning models that were trained in a supervised manner.

2.4 Quantization

As Transformer-based architectures continue to scale in size and capabilities, so do the memory and computational requirements. This introduces significant challenges when deploying these models for real-life applications, especially in resource constrained environments [25]. For this reason, many different approaches to model compression have been researched, including knowledge distillation (DeiT) [66], weight pruning [78], low-rank factorization [10] and quantization [50]. Among these techniques, quantization stands out due to its consistent success in both training and inference of AI models, and can significantly reduce the memory footprint and speed up model inference [27].

Quantization works by mapping high precision model weights to lower bit-widths without having to change the underlying model architecture. This is especially useful for Transformer-based architectures, which are often carefully designed. Furthermore, quantization also stimulates the development of more efficient hardware design utilizing lower-bit data [16, 37, 70].

Most quantization methods are applied post-training (PTQ), which does not require any changes to the existing training pipeline. PTQ directly quantizes the model weights without requiring re-training or fine-tuning, making it the current preferred choice for expensive large-scale models [49]. In practice, however, these methods also suffer from significant accuracy loss, especially for low-precision quantization (≤ 3 bits) [20, 25]. This performance degradation can be attributed to the significant difference between the original and quantized weights, for which the model is not optimized during training [38, 70].

Alternatively, Quantization-Aware Training (QAT) aims to alleviate this loss in accuracy by emulating the quantization during the training phase [25]. Models trained using QAT learn to adapt to the reduced precision right away, allowing these models to retain higher accuracy at inference-time. The main challenges of QAT lie in the gradient approximations for the non-differentiable quantization operations. Moreover, the increased performance of QAT at inference-time comes at the cost of being more time-consuming and resource intensive during training. This is because QAT requires retraining the model from scratch, and takes longer for lower precision models [20, 25].

In addition to lower memory consumption and faster inference, extreme low-level binary $\{-1, +1\}$ and ternary $\{-1, 0, +1\}$ quantization allows us to simplify matrix multiplications. Specifically, the expensive floating point multiplications can be replaced with efficient bitwise or addition operations, greatly improving the power efficiency of the model [79]. In this domain, Spiking Neural Networks quantize the activations [52, 80], whereas Binarized and Ternary Neural Networks (BNNs and TNNs) quantize the weights [2, 16].

For vision tasks, architectures such as BiViT [31] and BinaryViT [44] successfully applied

model binarization to Vision Transformers. However, while these models outperform previous binarization techniques, they still suffer from noticeable performance degradation.

In language modeling, BitNet demonstrated the scalability of 1-bit quantization in Transformer LLMs [70]. Specifically, the authors introduce the BitLinear layer, which combines 8-bit activation quantization with binary weights that are trained using QAT. BitNet is based on the succesful LLaMA architecture [68], only replacing all linear layers with BitLinear layers.

More recently, Ma et al. [51] introduced BitNet b1.58, replacing the binary weights with ternary weights. The authors show that this approach retains all the benefits of the original BitNet, while further increasing its computational and memory efficiency by introducing sparsity in the model. Additionally, the results show that BitNet b1.58 can match full precision baselines when scaled sufficiently. For this reason, exploring the BitLinear layer applied to modern ViT architectures could lead to more efficient and accessible models, while maintaining high classification accuracy.

3 | Methodology

In this chapter, we discuss our proposed architecture for this thesis. We first introduce our data preprocessing pipelines on the ImageNet-1k classification dataset [57], the FaceForensics++ [56] and Celeb-DF [48] deepfake detection datasets. Next, we discuss the base Vision Transformer architecture, as well as our proposed modifications, in Sec. 3.2. We replace the linear layer modules with our implementation of BitLinear as proposed by Wang et al. [51, 70], which we describe in detail in Sec. 3.3. Finally, in Sec. 3.4 we integrate knowledge distillation as proposed by Touvron et al. [66] to allow for more data-efficient training. The PyTorch [3] code implementation can be found on `https://github.com/ingur/` link

3.1 Data Preprocessing

We pretrain our custom Vision Transformer on ImageNet-1k [57]. The original ViT paper focused on large-scale pre-training on ImageNet-21k and JFT-300M [61], which require significant computational resources. However, recent advancements [45, 66] allow us to train more efficiently on mid-sized datasets, including ImageNet-1k, greatly reducing the cost of pre-training. After pre-training, we fine-tune our proposed model on two deepfake datasets, FaceForensics++ [56] and Celeb-DF [48]. These datasets are widely explored datasets, which makes it easier to compare architectural changes when training for these datasets [64]. Due to computational and resource constraints, we decided to only consider these two mid-size deepfake datasets, and leave third generation datasets (e.g. DFDC [23]) for future work.

ImageNet. The ImageNet-1k dataset was obtained through the HuggingFace [1] datasets library [46]. As this dataset contains multiple image modes, we first ensure all images follow a consistent *RGB* format using the `Image.convert()` method from the Pillow library. We also resize the images so that the smaller edges are always 256 pixels, while preserving the aspect ratio. This is done using the `torchvision.transforms.Resize()` method from PyTorch [3].

Deepfake datasets. The FaceForensics++ and Celeb-DF dataset are both available through their respective GitHub repositories [2] [3], and require access to be requested through a form. Both datasets consist of multiple directories of videos, and share a common preprocessing pipeline. We extract an equal number of frames from both the real and fake videos using a 30-frame interval. For FaceForensics++, we processed 1000 real and 1000 fake videos. For Celeb-DF, we used all 890 real and randomly sampled an equal amount of fake videos. Next, we use the MTCNN face detection model to detect faces in each frame, and discard frames without faces. We retain only the largest faces detected in each frame and ensure the model focuses on the most likely faces, and discard the remaining ones. Finally, each face is returned as a 256x256 cropped image, with 40-pixel margins.

[1] `https://huggingface.co/datasets/ILSVRC/imagenet-1k`
[2] `https://github.com/ondyari/FaceForensics`
[3] `https://github.com/yuezunli/celeb-deepfakeforensics`

Augmentation. Data augmentation is a popular method to reduce overfitting and improve generalization of deep learning models, done by generating additional training data [17]. Simple transformations such as horizontal flips and random cropping are commonly used in image classification or deepfake detection models [23, 43]. Our pre-training mainly relies on three recent augmentation techniques: RandAugment [17], MixUp [76] and CutMix [74].

RandAugment was introduced by Cubuk et al. [17] and can be used to apply a fixed number of random augmentation operations to images. Additionally, the strength and number of the applied augmentations can be controlled with two simple parameters. This results in a simple and consistent pipeline that has shown to outperform many complex augmentation strategies without extensive need of hyperparameter tuning.

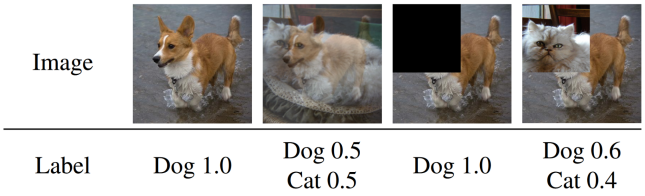

Figure 3.1: Comparison of various image augmentation techniques, from left to right: original image, Mixup [76], Cutout [21] and CutMix [74]. Figure from [74].

Next, MixUp and CutMix are used to create inter-class examples (see Fig. 3.1). MixUp constructs new training examples by taking linear interpolations of random pairs of images and their corresponding labels [76]. Given two training examples: (x_i, y_i) and (x_j, y_j), this method creates a new training example $(\tilde{x}, \tilde{y})$ using a mixing parameter λ from a beta distribution:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1 - \lambda) x_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j \end{aligned} \tag{3.1}$$

This method extends the training distribution and enhances model robustness and performance against adversarial examples.

Similarly, CutMix is another method that combines two image samples, and linearly interpolates the labels [74]. In contrast to CutOut [21], this method replaces a random image region with a patch from another training example, preventing inefficiency during training and leading to better localization capabilities.

3.2 Vision Transformer

This section describes our base Vision Transformer architecture. The proposed architecture builds upon the original ViT as proposed by Dosovitskiy et al. [24]. We apply minor simplifications and optimizations based on the work of various other research projects [6, 45, 60, 75]. We first describe the original ViT architecture, and then explain our modifications.

3.2.1 Patch embedding and positional encoding

The standard Transformer architecture is designed to receive a 1D sequence of D-dimensional token embeddings z_0 as input [69]. Let $x \in \mathbb{R}^{H \times W \times C}$ be a 2D input image[4], where (H, W)

[4]To simplify our formulations, we ignore the batch dimension.

denote the height and width, and C the number of channels. First, the input image is divided into non-overlapping flattened patch vectors:

$$\mathcal{P}(x) = [x_p^1, \ldots, x_p^N] \qquad x_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \tag{3.2}$$

Here, P indicates the patch size and $N = HW/P^2$ the resulting number of patches. Next, patch embeddings can be obtained through layer normalization (LN) [4] followed by a linear projection of each patch into the space of the Transformer encoder. This process is called tokenization ($\mathcal{T}$) and is defined as follows:

$$\mathcal{T}(x) = \mathrm{LN}(\mathcal{P}(x))\mathbf{E}_t = [x_p^1\mathbf{E}_t, \ldots, x_p^N\mathbf{E}_t] \qquad \mathbf{E}_t \in \mathbb{R}^{(P^2 \cdot C) \times D} \tag{3.3}$$

Here, $\mathbf{E}_t$ is the learnable embedding matrix for tokens, and D is the hidden dimensionality of the Transformer. Next, a special learnable classification token x_{class} is prepended to the embedded patches, which serves as a global representation of the entire image for classification tasks. Finally, learnable positional embeddings $\mathbf{E}_{pos}$ are added to retain positional information. The resulting sequence z_0 is then used as the input to the first Transformer encoder:

$$z_0 = [x_{\text{class}}, \mathcal{T}(x)] + \mathbf{E}_{pos} \qquad x_{\text{class}} \in \mathbb{R}^D,\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{3.4}$$

Shifted Patch Tokenization. We extend our tokenization by using the Shifted Patch Tokenization (SPT), which utilizes spatial relations between neighboring pixels during tokenization to capture more spatial information [45]. While originally introduced for ViT training on small-size datasets, SPT has also been found to improve ViTs even on mid-size datasets such as ImageNet-1k. SPT works by shifting the input image in four diagonal directions, creating four permutations of the input image. These permutations are then concatenated to the original image before being divided into non-overlapping patches and tokenized. This process is illustrated in Fig. 3.2 and is formulated as follows:

$$\mathrm{SPT}(x) = \mathrm{LN}(\mathcal{P}([x, x_s^1, \ldots, x_s^{N_S}]))\mathbf{E}_S \qquad \mathbf{E}_S \in \mathbb{R}^{(P^2 \cdot C \cdot (N_S+1)) \times D} \tag{3.5}$$

where x_s^i represents the i-th shifted image, N_S the number of image shifts and $\mathbf{E}_S$ the new learnable linear projection.

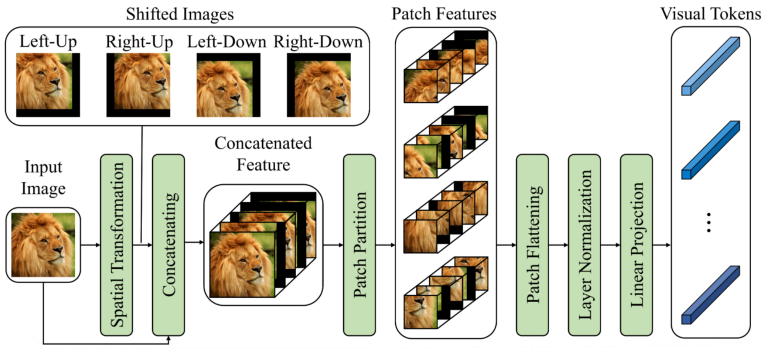

Figure 3.2: Overview of Shifted Patch Tokenization (SPT). Shifted permutations of the input image are concatenated before being transformed to patch embeddings (visual tokens). Figure from [45].

Positional Encoding. The Vision Transformer architecture as introduced by Dosovitskiy et al. [24] uses learnable positional embeddings $\mathbf{E}_{pos}$. We replace these embeddings with fixed sinusoidal 2D positional encodings to reduce the amount of trainable parameters. This approach extends the 1D sinusoidal encodings introduced in the original Transformer architecture [69] to 2D, and provides a simple representation of position comparable to learnable embeddings [12]. We first create a 2D grid of (x, y) coordinate pairs for each patch. The sinusoidal positional encoding can then be computed for each coordinate pair and dimension i ranging from 0 to $D/4 - 1$, resulting in a unique D-dimensional encoding for each patch:

$$\begin{aligned} E_{\text{sin-cos}}(x, y, 4i) &= \sin(x \cdot \omega_i) \\ E_{\text{sin-cos}}(x, y, 4i + 1) &= \cos(x \cdot \omega_i) \\ E_{\text{sin-cos}}(x, y, 4i + 2) &= \sin(y \cdot \omega_i) \\ E_{\text{sin-cos}}(x, y, 4i + 3) &= \cos(y \cdot \omega_i) \\ \omega_i &= 1/10000^{4i/D} \end{aligned} \tag{3.6}$$

Note that for this encoding, the dimensionality D must be divisible by 4.

Global average-pooling. Instead of using a learnable classification token x_{class}, we use global average-pooling (GAP) to obtain the global representation of the image. Chen et al. [12] show that using GAP yields comparable performance to a classification token. Furthermore, the computational cost is reduced as GAP does not require an extra token to be processed through all attention layers, increasing the model efficiency. We calculate the global representation by averaging the activations across all patch embeddings from the final Transformer layer. This operation can be formulated as follows:

$$z_{\text{GAP}} = \frac{1}{N} \sum_{i=1}^{N} z_L^i \qquad z_L \in \mathbb{R}^{N \times D},\ z_{\text{GAP}} \in \mathbb{R}^D \tag{3.7}$$

where z_L^i represents the i-th token embedding from the output of the final Transformer layer L, and N the number of patch embeddings.

Our modified input to the Transformer encoder can now be defined as:

$$z_0 = \text{SPT}(x) + E_{\text{sin-cos}} \qquad z_0 \in \mathbb{R}^{N \times D} \tag{3.8}$$

We combine Shifted Patch Tokenization with 2D sinusoidal positional encoding as input to our Transformer, and use global average-pooling to capture the final image representation as opposed to a classification token.

3.2.2 Transformer encoder

The Transformer encoder consists of L identical layers, and consists of two components: the multi-head self-attention mechanism (MHA) and a fully-connected Feed Forward Network (FFN) [69]. Layer normalization (LN) is applied before each component [4], and residual connections are used after each block [30]. This architecture can be formulated as follows:

$$z'_l = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1} \qquad l = 1 \dots L \tag{3.9}$$

$$z_l = \text{FFN}(\text{LN}(z'_l)) + z'_l \qquad l = 1 \dots L \tag{3.10}$$

Here, $z_l \in \mathbb{R}^{N \times D}$ is the output of layer l, and z_0 is the input to the first layer as defined in Eq. (3.8).

The FFN block uses two linear projections, and a GELU activation function [32]:

$$\text{FFN}(z) = \text{GELU}(z\mathbf{E}_1)\mathbf{E}_2 \qquad \mathbf{E}_1 \in \mathbb{R}^{D \times D_{hd}}, \mathbf{E}_2 \in \mathbb{R}^{D_{hd} \times D} \tag{3.11}$$

$$\text{GELU}(x) = x\Phi(x) \tag{3.12}$$

where $z \in \mathbb{R}^{N \times D}$ is the input embeddings, $\mathbf{E}_1$ and $\mathbf{E}_2$ the learnable projections, D_{hd} the hidden dimensionality of the FFN, and $\Phi(x)$ the standard Gaussian cumulative distribution function.

The self-attention mechanism works by first creating three matrices: the Query (Q), Key (K) and Value matrix (V). These matrices are obtained through a learnable linear projection ($\mathbf{E}^{QKV}$) of the input:

$$[Q, K, V] = z\mathbf{E}^{QKV} \qquad \mathbf{E}^{QKV} \in \mathbb{R}^{D \times 3D_h},\ Q, K, V \in \mathbb{R}^{N \times D_h} \tag{3.13}$$

where D_h corresponds to the dimension of each attention head. Next, the self-attention (SA) weights are computed as the scaled dot product between the query with all keys, followed by a softmax operation and multiplication with the Value matrix:

$$\text{SA}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{D_h}})V \tag{3.14}$$

The scaling factor $\sqrt{D_h}$ is used to prevent the softmax function from having small gradients. For multi-head self-attention (MSA), SA is extended to use multiple self-attention heads in parallel:

$$\begin{aligned} \text{MSA}(z) &= [\text{head}_1(z), \ldots, \text{head}_h(z)]\mathbf{E}^O & \mathbf{E}^O &\in \mathbb{R}^{h \cdot D_h \times D} \\ \text{head}_i(z) &= \text{SA}(z\mathbf{E}_i^Q, z\mathbf{E}_i^K, z\mathbf{E}_i^V) & \mathbf{E}_i^Q, \mathbf{E}_i^K, \mathbf{E}_i^V &\in \mathbb{R}^{D \times D_h} \end{aligned} \tag{3.15}$$

Here, h is the number of attention heads and $\mathbf{E}_i^Q, \mathbf{E}_i^K$ and $\mathbf{E}_i^V$ are the learnable projections for the i-th head. The outputs of the individual heads is concatenated, and $\mathbf{E}^O$ is used to project the concatenated outputs back to the encoder dimensionality D. Moreover, D_h is usually set to D/h, keeping the computational cost of MSA similar to single-head attention with full dimensionality. Unfortunately, the computational complexity of self-attention scales quadratically $O(N^2)$ with the number of input tokens. And while various sub-quadratic approaches to self-attention have been proposed [18, 41], these approaches often come with different trade-offs in model performance. For this reason, we focus on standard self-attention in this thesis.

Dropout. The original ViT architecture uses dropout [59] in their larger models [24]. Dropout is a regularization technique that randomly sets the activations of some neurons in a layer to 0 at each update during training time. This reduces co-adaption between neurons, which can greatly reduce the effects of overfitting. However, recent studies [6, 60] suggest that data augmentation is more effective than regularization techniques like dropout and stochastic depth [36], especially for smaller ViT models. Therefore, we exclude regularization such as dropout and stochastic depth for this thesis.

Layer Normalization. Layer Normalization (LN) is a technique introduced by Ba et al. [4] which helps stabilize training, allowing for faster convergence. Given an input vector $x \in \mathbb{R}^n$, Layer normalization works by re-centering and re-scaling the activations as follows:

$$\mathrm{LN}(x) = \frac{x - \mu}{\sigma} \cdot \gamma + \beta \qquad \gamma, \beta \in \mathbb{R}^n \tag{3.16}$$

where γ and β are learnable gain and bias parameters, respectively. The mean μ and standard deviation σ are computed as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3.17}$$

While layer normalization has been successfully applied to the ViT architecture [24], it also introduces computational overhead due to the extra statistics (μ, σ) that have to be calculated. To improve the efficiency of our architecture, we replace Layer Normalization with Root Mean Square Layer Normalization (RMSNorm) [75]. RMSNorm is an efficient alternative to Layer Normalization that removes the mean statistic, and has been successfully applied to Transformer LLMs like LLaMA [68]. The authors of RMSNorm argue that the re-centering of the activations contributes little to the success of layer normalization, and focus only on re-scaling the invariance and regularizing according to the root-mean-square (RMS) statistic:

$$\mathrm{RMSNorm}(x) = \frac{x}{\mathrm{RMS}(x)} \cdot \gamma \tag{3.18}$$

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{3.19}$$

The resulting method is used as a drop-in replacement for all instances of layer normalization in our architecture, and further increases the efficiency of our model.

3.3 BitLinear layer

In this section, we describe our implementation of the BitLinear layer as introduced by Wang et al. [70] for their BitNet LLM architecture. Specifically, we implement the BitLinear module proposed for BitNet b1.58 [51], which uses ternary weights $\{-1, 0, +1\}$ rather than the original binarized $\{-1, +1\}$ architecture. Our PyTorch [3] implementation can be found on GitHub: `https://github.com/ingur/bitlinear-pytorch`.

The matrix multiplication in a standard full-precision linear layer can be expressed as follows:

$$y_i = (x\mathbf{W})_i = \sum_{j=1}^{n} x_j \mathbf{W}_{ij} \qquad x \in \mathbb{R}^{1\times n}, \mathbf{W} \in \mathbb{R}^{n\times m}, y \in \mathbb{R}^{1\times m} \tag{3.20}$$

where x is the input, $\mathbf{W}$ is the weight matrix and y the output. Using ternary weights constrained to the set $\{-1, 0, +1\}$, we can now replace the multiplication operation with pure addition operations. Let $\tilde{\mathbf{W}} \in \mathbb{R}^{n\times m}$ be the ternary weight matrix, where for each $\tilde{\mathbf{W}}_{ij} \in \{-1, 0, +1\}$. The multiplication operations can now be replaced by efficient addition or subtraction operations:

$$x_j \tilde{\mathbf{W}}_{ij} = \begin{cases} x_j, & \text{if } \tilde{\mathbf{W}}_{ij} = 1, \\ 0, & \text{if } \tilde{\mathbf{W}}_{ij} = 0, \\ -x_j, & \text{if } \tilde{\mathbf{W}}_{ij} = -1, \end{cases} \tag{3.21}$$

Which in turn allows us to express ternary matrix multiplication as:

$$\tilde{y}_i = \sum_{j=1}^{n} x_j \tilde{\mathbf{W}}_{ij} = \sum_{j:\tilde{\mathbf{W}}_{ij}=1}^{n} x_j - \sum_{j:\tilde{\mathbf{W}}_{ij}=-1}^{n} x_j \tag{3.22}$$

with $\tilde{y} \in \mathbb{R}^{1\times m}$ as the new output. Note that this also works for binary weights constrained to the set $\{-1, +1\}$, as the 0 weight only acts as a feature filter that introduces sparsity. The weight quantization uses an absmean quantization function, where we scale the weight matrix by its average absolute value and round each value to its nearest integer in the range $\{-1, 0, +1\}$:

$$\tilde{\mathbf{W}} = \text{RoundClip}(\frac{\mathbf{W}}{\beta + \epsilon}, -1, 1) \tag{3.23}$$

$$\text{RoundClip}(x, a, b) = \max(a, \min(b, \text{round}(x))) \tag{3.24}$$

$$\beta = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |W_{ij}| \tag{3.25}$$

where ϵ is a small floating point number to avoid division by zero errors. The activations are also quantized to 8-bit precision using absmax quantization [20] as defined in Eq. (3.26). Absmax quantization scales the activations into 8-bit range $[-128, 127]$ by multiplying the activations with 127 and dividing by the absolute maximum activation value:

$$\tilde{x} = \text{ActQuant}(x) = \text{RoundClip}(x \cdot \gamma, -128, 127) \tag{3.26}$$

$$\gamma = \frac{127}{\max(|x|, \epsilon)} \tag{3.27}$$

In full-precision linear layers, initialization methods such as Kaiming or Xavier initialization help maintain numerical stability throughout training [28, 29]. However, the non-linearity of the activation quantization does not preserve this variance. For this reason, RMSNorm (see Eq. (3.18)) is used before the activation quantization to help maintain numerical stability and preserve the variance. We ensure that RMSNorm is used before any linear layer throughout the model, including the patch tokenization process (SPT), before attention (MHA) and in the feedforward networks (FFN). Finally, the output activations are rescaled using β (Eq. (3.25)) and γ (Eq. (3.27)) to dequantize them to their original precision. The BitLinear layer can now be defined as follows:

$$y = \tilde{x}\tilde{\mathbf{W}} = \text{ActQuant}(\text{RMSNorm}(x))\tilde{\mathbf{W}} \times \frac{\beta}{\gamma} \tag{3.28}$$

This approach allows for future optimized hardware and kernel implementations to take advantage of the ternary weights for faster and more efficient matrix multiplication. However, for our

training and experiment purposes, we apply the scaling factors before using the `F.Linear()` operation in PyTorch.

Straight-through estimator. Backpropagation requires all functions in our network to be differentiable. However, the round and clip functions used in the weight and activation quantization (see Eq. (3.24)) are non-differentiable. To address this issue, Wang et al. [70] use straight-through estimators (STEs) to allow the gradients to flow through these non-differentiable functions [5]. During training, we quantize the weights during each forward pass ($\tilde{\mathbf{W}}$). However, during the backward pass, the gradients are calculated with respect to the high-precision latent weights ($\mathbf{W}$).

3.4 Knowledge Distillation

Hinton et al. [34] introduced Knowledge Distillation as a training paradigm utilizing soft-labels from a strong teacher model to efficiently train smaller student models. Here, soft-labels refer to the raw logits as produced by the teacher network, as opposed to the maximum of scores, which gives a hard-label. Moreover, knowledge distillation can also be used to compress the knowledge of larger models into smaller models. Touvron et al. [66] expanded upon their work by introducing a new distillation strategy designed for ViT models, resulting in the Data-Efficient Image Transformer (DeiT). Using this strategy on ImageNet-1k, DeiT is able to outperform the original ViT-B model pretrained on the JFT-300M dataset [61].

Based on the results of Touvron et al. [66], we implement hard-label distillation due to its increased performance over soft-label distillation. Let y be the ground-truth label, and $y_t = \text{argmax}_c z_t(c)$ the one-hot encoded hard decision of the teacher model. We append a learnable distillation token to the initial patch embeddings z_0, which interacts with the other embeddings through self-attention, similarly to a class token x_{class}. Let z_s represent the final output of our vision transformer, calculated using global-average pooling (see Eq. (3.7)), and z_d the final output distillation token. The hard-label distillation loss can now be defined as:

$$\mathcal{L}^{\text{hard}}_{\text{global}} = (1-\alpha)\mathcal{L}_{\text{CE}}(\psi(z_s), y) + \alpha\mathcal{L}_{\text{CE}}(\psi(z_d), y_t) \tag{3.29}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy function, and ψ the softmax function. Additionally, α controls the tradeoff between main loss and distillation loss, and is set to 0.5 by default. This approach allows our ViT model to efficiently learn from a strong teacher classifier model, further improving its performance.

4 | Experiments

5 | Results

6 | Conclusion

Bibliography

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, December 2018. doi: 10.1109/wifs.2018.8630761.

[2] Hande Alemdar, Vincent Leroy, Adrien Prost-Boucle, and Frédéric Pétrot. Ternary Neural Networks for Resource-Efficient AI Applications. *arXiv:1609.00222*, 2017.

[3] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL `https://pytorch.org/assets/pytorch2-2.pdf`.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

[5] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *CoRR*, abs/1308.3432, 2013. URL `http://arxiv.org/abs/1308.3432`.

[6] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain ViT baselines for ImageNet-1k. *arXiv:2205.01580*, 2022.

[7] Omkar Bhilare, Rahul Singh, Vedant Paranjape, Sravan Chittupalli, Shraddha Suratkar, and Faruk Kazi. Deepfake cli: Accelerated deepfake detection using fpgas, 2022. URL `https://arxiv.org/abs/2210.14743`.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018. URL `https://arxiv.org/abs/1710.08092`.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL `https://arxiv.org/abs/2104.14294`.

[10] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention approximation, 2021. URL https://arxiv.org/abs/2110.15343.

[11] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.02057.

[13] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy. *Calif. L. Rev.. California Law Review*, 107(IR):1753, 2019. URL http://lawcat.berkeley.edu/record/1136469.

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. URL https://arxiv.org/abs/1610.02357.

[15] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. *Combining EfficientNet and Vision Transformers for Video Deepfake Detection*, page 219–229. Springer International Publishing, 2022. ISBN 9783031064333. URL http://dx.doi.org/10.1007/978-3-031-06433-3_19.

[16] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. *arXiv:1602.02830*, 2016.

[17] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019. URL https://arxiv.org/abs/1909.13719.

[18] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

[19] Sowmen Das, Selim Seferbekov, Arup Datta, Md. Saiful Islam, and Md. Ruhul Amin. Towards solving the deepfake problem : An analysis on improving deepfake detection using dynamic face augmentation, 2021. URL https://arxiv.org/abs/2102.09603.

[20] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv:2208.07339*, 2022.

[21] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. URL https://arxiv.org/abs/1708.04552.

[22] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset, 2019. URL https://arxiv.org/abs/1910.08854.

[23] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*, 2020.

[25] Dayou Du, Gu Gong, and Xiaowen Chu. Model quantization and hardware acceleration for vision transformers: A comprehensive survey, 2024. URL https://arxiv.org/abs/2405.00314.

[26] N. Fufour, P. Gully, A. Karlsson, A.V. vorbyov, T. Leung, J. Childs, and C. Bregler. Deepfakes detection dataset by google & jigsaw, 2019.

[27] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference, 2021. URL https://arxiv.org/abs/2103.13630.

[28] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. URL https://arxiv.org/abs/1502.01852.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[31] Yefei He, Zhenyu Lou, Luoming Zhang, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Bivit: Extremely compressed binary vision transformer, 2023. URL https://arxiv.org/abs/2211.07091.

[32] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

[33] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. Deepfake detection scheme based on vision transformer and distillation, 2021. URL https://arxiv.org/abs/2104.01353.

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

[35] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, 2019. URL https://arxiv.org/abs/1901.09891.

[36] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth, 2016. URL https://arxiv.org/abs/1603.09382.

[37] Mingqiang Huang, Junyi Luo, Chenchen Ding, Zikun Wei, Sixiao Huang, and Hao Yu. An integer-only and group-vector systolic accelerator for efficiently mapping vision transformer on edge. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(12):5289–5301, 2023. doi: 10.1109/TCSI.2023.3312775.

[38] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms, 2024. URL https://arxiv.org/abs/2402.04291.

[39] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, 2020. URL https://arxiv.org/abs/2001.03024.

[40] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

[41] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020. URL https://arxiv.org/abs/2006.16236.

[42] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018. URL https://arxiv.org/abs/1812.08685.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[44] Phuoc-Hoan Charles Le and Xinlin Li. Binaryvit: Pushing binary vision transformers towards convolutional models, 2023. URL https://arxiv.org/abs/2306.16678.

[45] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv:2112.13492*, 2021.

[46] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

[47] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts, 2019.

[48] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *CoRR*, abs/1909.12962, 2019. URL http://arxiv.org/abs/1909.12962.

[49] Zhikai Li, Xuewen Liu, Jing Zhang, and Qingyi Gu. Repquant: Towards accurate post-training quantization of large transformer models via scale reparameterization, 2024. URL https://arxiv.org/abs/2402.05628.

[50] Zhenhua Liu, Yunhe Wang, Kai Han, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer, 2021. URL https://arxiv.org/abs/2106.14156.

[51] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. *arXiv:2402.17764*, 2024.

[52] Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. ISSN 0893-6080. URL https://www.sciencedirect.com/science/article/pii/S0893608097000117.

[53] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis, 2024. URL https://arxiv.org/abs/2405.00355.

[54] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019.

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[56] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019. URL http://arxiv.org/abs/1901.08971.

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

[59] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[60] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2022. URL https://arxiv.org/abs/2106.10270.

[61] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017. URL https://arxiv.org/abs/1707.02968.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[63] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

[64] Vrizlynn L. L. Thing. Deepfake detection with deep learning: Convolutional neural networks versus transformers, 2023. URL https://arxiv.org/abs/2304.03698.

[65] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance, 2020. URL https://arxiv.org/abs/2004.07532.

[66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv:2012.12877*, 2021.

[67] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021. URL https://arxiv.org/abs/2103.17239.

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[70] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. BitNet: Scaling 1-bit Transformers for Large Language Models. *arXiv:2310.11453*, 2023.

[71] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, 2020. URL https://arxiv.org/abs/1909.06122.

[72] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer, 2021. URL https://arxiv.org/abs/2102.11126.

[73] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses, 2018. URL https://arxiv.org/abs/1811.00661.

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. URL https://arxiv.org/abs/1905.04899.

[75] Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL https://arxiv.org/abs/1910.07467.

[76] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. URL https://arxiv.org/abs/1710.09412.

[77] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. ISSN 1558-2361. URL http://dx.doi.org/10.1109/LSP.2016.2603342.

[78] Mingjian Zhu, Yehui Tang, and Kai Han. Vision transformer pruning, 2021. URL https://arxiv.org/abs/2104.08500.

[79] Rui-Jie Zhu, Yu Zhang, Ethan Sifferman, Tyler Sheaves, Yiqiao Wang, Dustin Richmond, Peng Zhou, and Jason K. Eshraghian. Scalable matmul-free language modeling, 2024. URL https://arxiv.org/abs/2406.02528.

[80] Rui-Jie Zhu, Qihang Zhao, Guoqi Li, and Jason K. Eshraghian. Spikegpt: Generative pre-trained language model with spiking neural networks, 2024. URL https://arxiv.org/abs/2302.13939.