Literature Review for Ingur Thesis References

Reza

August 19, 2024

BinaryViT [1]

0.1 Binarization in Neural Networks

0.1.1 Convolutional Neural Networks (CNNs):

Binarization in CNNs has been extensively explored. The idea of using binary weights and activations in CNNs to reduce memory and computation dates back to works like XNOR-Net (Rastegari et al., 2016). These works showed that CNNs could achieve competitive performance while significantly reducing resource consumption.

Binary Neural Networks (BNNs):

values for both weights and activations, leading to models that require less storage and computational power.

Highlight how CNN architectures, especially with their convolutional nature, lend themselves to being binarized without extreme performance degradation.

0.1.2 Vision Transformers (ViTs):

ViTs, since being introduced by Dosovitskiy et al. (2020), have demonstrated significant improvements over CNNs in various vision tasks by modeling long-range dependencies through self-attention mechanisms.

ViTs, however, face challenges related to computational cost, making them difficult to deploy on edge devices.

In comparison to CNNs, few studies have explored binarizing ViTs due to their unique architecture. Some early attempts at transformer binarization (such as BiMLP) explore this to a limited degree but focus mainly on feed-forward layers and do not thoroughly analyze architectural impacts.

0.2 Existing Challenges in Binarizing ViTs

1. Unlike CNNs, directly applying existing binarization techniques to ViTs often results in significant performance degradation, particularly on large datasets like ImageNet-1k. CNNs benefit from convolutional layers that

preserve spatial information, whereas ViTs' token-based structure lacks this property, leading to lower representational capability when binarized.

2. Mention that CNN-based binarization techniques (such as ReBNet and XNOR-Net) have been successful because they account for the architectural benefits of convolutions that help preserve important spatial information even in binary form.

3. Summarize existing research like BinaryViT, which seeks to overcome these limitations by introducing architectural modifications from CNNs into a ViT framework.

0.3 BinaryViT and its Novel Contributions

In [1] proposes a new way to introduce architectural innovations inspired by CNNs into a binary ViT model, without actually introducing convolutions. Key contributions of this model include:

1. The use of average pooling layers instead of token pooling layers.

2. The introduction of multiple pooling branches to help compensate for the loss of representational power.

3. Affine transformations and the addition of a pyramid structure that allows binary features to be processed at higher resolutions while minimizing computational complexity.

Discuss how BinaryViT leverages elements from CNNs while maintaining the flexibility of ViT architectures. Explain how your approach addresses the shortcomings of previous binarization efforts for ViTs by preserving essential architectural properties.

Bibliography

[1] P.-H. C. Le and X. Li, “Binaryvit: pushing binary vision transformers towards convolutional models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4664–4673.