BINARIZED NEURAL NETWORKS FOR RESOURCE-CONSTRAINED ON-DEVICE GAIT IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

User authentication through gait analysis is a promising application of discriminative neural networks – particularly due to the ubiquity of the primary sources of gait acceleometry, in-pocket cellphones. However, conventional machine learning models are often too large and computationally expensive to enable inference on low-resource mobile devices. We propose that binarized neural networks can act as robust discriminators, maintaining both an acceptable level of accuracy while also dramatically decreasing memory requirements, thereby enabling on-device inference. To this end, we propose BiPedalNet, a compact CNN that nearly matches the state-of-the-art on the Padova gait dataset, with only **1/32** of the memory overhead.

1 INTRODUCTION

1.1 GAIT IDENTIFICATION

Human gait, as measured via smartphone sensors, has been shown to be a viable biometric for distinguishing users with high accuracy (Derawi et al. (2010)). Both user privacy and scalability concerns suggest re-identification is best done on-device. However, previous work on user re-identification from gait data consistently demonstrates high accuracy with deep CNNs, but also high computational overhead, making on-device, real-time, inference, intractable (Gadaleta & Rossi (2018)).

1.2 BINARIZED NEURAL NETWORKS

Binarized Neural Networks (Courbariaux et al. (2016)) are neural networks with weights constrained to $\{-1, 1\}$. BNNs, although generally achieving lower accuracies than their full-precision equivalents, are both smaller, due to compact binary weight matrices, and faster, due to the usage of bitwise operations for matrix multiplication and activation functions.

This makes BNNs an ideal choice for on-device machine learning. Mobile devices, particularly budget devices, often have strict limitations on both memory availability and processing power, making frugality an important virtue. We propose binarized neural networks as a computationally inexpensive way to conduct lightweight gait identification.

2 METHODOLOGY

2.1 TRAINING BINARIZED WEIGHTS

In order to train the binarized weights of the architecture, we maintain latent real-valued weights proposed in Courbariaux et al. (2016) that are updated during backpropagation and binarized during the forward pass. These latent weights are set to 1 if positive; otherwise, clipped to -1 in the forward pass. Note that these latent weights are not used at test time, and thus don't contribute to the memory overhead.

2.2 ARCHITECTURE

We propose a lightweight, binarized neural network architecture, dubbed BiPedalNet 1. BiPedalNet is built primarily using convolutional neural layers. Having 2-dimensional convolutions early in

Layer	Units	Filter Size	# 1-bit params	# 32-bit params
Conv2D	32	(3,3)	288	0
MaxPool2D	-	(2,2)	0	0
BatchNorm	-	-	0	96
Conv1D	64	(1,3)	6144	0
BatchNorm	-	-	0	192
Conv1D	64	(1,3)	12288	0
BatchNorm	-	-	0	192
Flatten	-	-	-	-
Dense	32	-	276480	0
BatchNorm	-	-	0	96
Dense	38	-	1216	0
Softmax	-	-	-	-

Table 1: The BiPedalNet Architecture. All layers have binary weights.

the network allows the model to learn cross-channel dependencies between the accelerometric and gyroscopic dimensions. Once the cross-channel features are extracted, we restrict the convolutions to a single dimension to confine learning to the temporal dimension. For regularization and pooling, we found the combination of max pooling and batch normalization yields the best performance. The binarized architecture is implemented with the LARQ (Geiger et al. (2019)) library.

2.3 Training Procedure

To compare our architecture against the existing state-of-the-art, we train and validate our architecture on the Padova gait dataset Gadaleta & Rossi (2018) (80/20 split). We compare BiPedalNet against a binzarized version of the IDNet model proposed in Gadaleta & Rossi (2018), with both models trained using the same hardware (1x Tesla V100 GPU).

We use the Adam optimizer (Kingma & Ba (2014)) to update the latent weights of the architecture, with a default learning rate of 1e-3 and a scheduler that exponentially reduces the rate when validation accuracy plateaus. The multi-class cross-entropy metric is used as the optimization objective.

3 Results and Discussion

Model	Number of Parameters	Size on disk (MB)	Top-1 Accuracy(%)
IDNet	335k	1.28	98.05
Binarized IDNet	335k	0.04	41.45
BiPedalNet	297k	0.04	**95.91**

Table 2: Quantitative Results for performance on Padova dataset

The first two results in Table 2 demonstrate that state-of-the-art performance on gait data does not persist through off-the-shelf binarization of full-precision networks. To achieve performance comparable to the full-precision model, special care needs to be applied in tuning the architecture to both the dataset and the task at hand. This is demonstrated by the superior performance of BiPedalNet over the binarized IDNet variant.

4 Conclusion

We've shown that binarized neural networks can achieve comparable accuracies to full precision architectures, at a fraction of the size, on gait identification tasks. However, it is clear that custom-designing binarized architectures is necessary for reasonable performance – naively binarizing a given full precision architecture is a recipe for disaster. Binarized neural networks offer a promising avenue to achieve low latency, low memory, inference in hardware-constrained devices, and we expect these models to perform well in low-resource mobile devices.

REFERENCES

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 preprint arXiv:1602.02830*, 2016.

Mohammad Omar Derawi, Claudia Nickel, Patrick Bours, and Christoph Busch. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 306–311. IEEE, 2010.

Matteo Gadaleta and Michele Rossi. Idnet: Smartphone-based gait recognition with convolutional neural networks. *Pattern Recognition*, 74:25 – 37, 2018. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.09.005. URL http://www.sciencedirect.com/science/article/pii/S0031320317303485.

Lukas Geiger, James Widdicombe, Arash Bakhtiari, Koen Helwegen, Maria Heuss, and Roeland Nusselder. Larq: An open-source deep learning library for training binarized neural networks. Web page, 2019. URL https://larq.dev.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

5 Responses to reviewer comments

5.1 Reviewer One:

Figure 1 doesn't convey much information. I would switch it out for something else.

We've replaced the original Figure 1 with a diagram of our proposed architecture.

1) Discuss why this new architecture or the use of the 'latent' weights provides improved gains.

We've added an additional explanation of the intuition behind why the new, latent-weight architecture provides these gains in our methodology section.

2) Test on multiple datasets to make sure that the improvements are not tied to the gait dataset that you tested.

We think this is beyond the scope of this work – plenty of literature exists on binarized neural networks, with benchmarks on a wide variety of tasks; our goal in this work was simply to demonstrate the viability of BNNs for the specific task of gait identification for low-resource devices, a setting commonly found when applying machine learning to the developing world.

5.2 Reviewer Two:

The paper lacks an in depth description of the methodology, design choices for the neural network architecture, and justification/ablation studies for these decisions.

Furthermore, the authors did not include relevant details on the training procedure or data used to obtain these results. Without these details, the results presented can not be reproduced or validated.

We've added additional details behind the training procedure and architecture design choices to our methodology, and have also linked to the dataset.

A comparison to other methods for model compression such as pruning or different quantization regimes would strengthen the the case for their proposed procedure.

Similarly to our response to Reviewer One's last comment, there already exist both theoretical comparisons and empirical results for binarized neural networks against other quantization regimes. We decided that these ablation experiments were outside the scope of a two-page submission.