Supplementary Material for Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources

Adrian Bulat and Georgios Tzimiropoulos
Computer Vision Laboratory, The University of Nottingham
Nottingham, United Kingdom
{adrian.bulat, yorgos.tzimiropoulos}@nottingham.ac.uk

1. Additional ablation studies

This section provides additional details for some of the ablation studies reported in Section 6.

Pooling type. In the context of binary networks, and because the output is restricted to 1 and -1, max-pooling might result in outputs full of 1s only. To limit this effect, we placed the activation function before the convolutional layers as proposed in [5, 9]. Additionally, we opted to replace max-pooling with average pooling. However, this leads to slightly worse results (see Table 1). In practice, we found that the use of blocks with pre-activation suffices and that the ratio of 1 and -1 is close to 50% even after max-pooling.

Layer type	# parameters	PCKh
(Ours, Final) + Average	6.2M	71.9%
(Ours, Final) + Max	**6.2M**	**76%**

Table 1: The effect of using different pooling methods when training our binary network in terms of PCKh-based performance on MPII validation set.

With or without ReLU. Because during the binarization process all ReLU layers are replaced with the Sign function, one might wonder if ReLUs are still useful for the binary case. Our findings are in line with the ones reported in [9]. By adding a ReLU activation after each convolutional layer, we observe a 2% performance improvement (see Table 2), which can be attributed to the added non-linearity, particularly useful for training very deep architectures.

Layer type	# parameters	PCKh
(Ours, Final)	6.2M	76%
(Ours, Final) + ReLU	**6.2M**	**77.8%**

Table 2: The effect of using ReLU when training our binary network in terms of PCKh-based performance on MPII validation set.

Performance. In theory, by replacing all floating-point multiplications with bitwise XOR and making use of the SWAR (Single instruction, multiple data within a register) [9, 4], the number of operations can be reduced up to 32x when compared against the multiplication-based convolution. However, in our tests, we observed speedups of up to 3.5x, when compared against cuBLAS, for matrix multiplications, a result being in accordance with those reported in [4]. As GPUs are already available on mobile devices, we did not conduct experiments on CPUs. However, given the fact that we used the same method for binarization as in [9], similar improvements in terms of speed, of the order of 58x, are to be expected: as the real-valued network takes 0.67 seconds to do a forward pass on a i7-3820 using a single core, a speedup close to x58 will allow the system to run in real-time.

In terms of memory compression, by removing the biases, which have minimum impact (or no impact at all) on performance, and by grouping and storing every 32 weights in one variable, we can achieve a compression rate of 39x when compared against the single precision counterpart of Torch. See also Fig. 1.

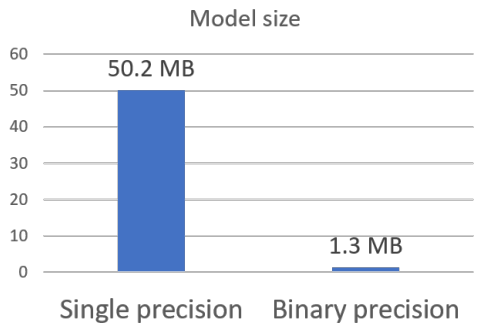


Figure 1: Memory compression ratio. By binarizing the weights and removing the biases, we achieve a compression rate of 39x when compared against the single precision model.

2. Additional face alignment results

This section provides additional numerical results on AFLW-PIFA and AFLW2000-3D.

PIFA [6]	RCPR [2]	PAWF [7]	CALE [1]	Ours
8.04	6.26	4.72	**2.96**	3.02

Table 3: NME-based (%) comparison on AFLW-PIFA evaluated on visible landmarks only. The results for PIFA, RCPR and PAWF are taken from [7].

CALE [1]	Ours
4.97	**4.47**

Table 4: NME-based (%) based comparison on AFLW-PIFA evaluated on all 34 points, both visible and occluded.

Method	[0,30]	[30,60]	[60,90]	Mean
RCPR(300W) [2]	4.16	9.88	22.58	12.21
RCPR(300W-LP) [2]	4.26	5.96	13.18	7.80
ESR(300W) [3]	4.38	10.47	20.31	11.72
ESR(300W-LP) [3]	4.60	6.70	12.67	7.99
SDM(300W) [10]	3.56	7.08	17.48	9.37
SDM(300W-LP) [10]	3.67	4.94	9.76	6.12
3DDFA [11]	3.78	4.54	7.93	5.42
3DDFA+SDM [11]	3.43	4.24	7.17	4.94
Ours	**2.47**	**3.01**	**4.31**	**3.26**

Table 5: NME-based (%) based comparison on AFLW2000-3D evaluated on all 68 points, both visible and occluded. The results for RCPR, ESR and SDM are taken from [11].

3. Facial part segmentation experiment

To show that the proposed block generalizes well, producing consistent results across various datasets and tasks, in this section, we report the results of an experiment on semantic facial part segmentation. To this end, we constructed a dataset for facial part segmentation by joining together the 68 ground truth keypoints (originally provided for face alignment) to fully enclose each facial component. In total, we created seven classes: skin, lower lip, upper lip, inner mouth, eyes, nose and background. Fig. 2 shows an example of a ground truth mask.

In particular, we trained the network on the 300W dataset (approx. 3000 images) and tested it on the 300W competition testset, both Indoor&Outdoor subsets (600 images), using the same procedure described in Section 7.

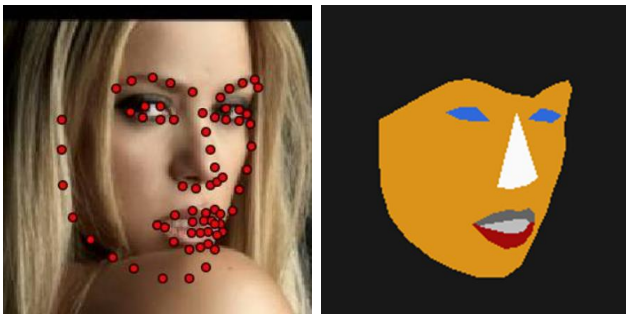


Figure 2: Example of a ground truth mask (right) produced by joining the 68 ground truth keypoints (left). Each color denotes one of the seven classes.

Architecture. We reused the same architecture for landmark localization, changing only the last layer in order to accommodate the different number of output channels (from 68 to 7). We report results for three different networks of interest: (a) a real-valued network using the original bottleneck block (called "Real, Bottleneck"), (b) a binary network using the original bottleneck block (called "Binary, Bottleneck"), and (c) a binary network using the proposed block (called "Binary, Ours"). To allow for a fair comparison, all networks have a similar number of parameters and depth. For training the networks, we used the LogSoftmax loss [8].

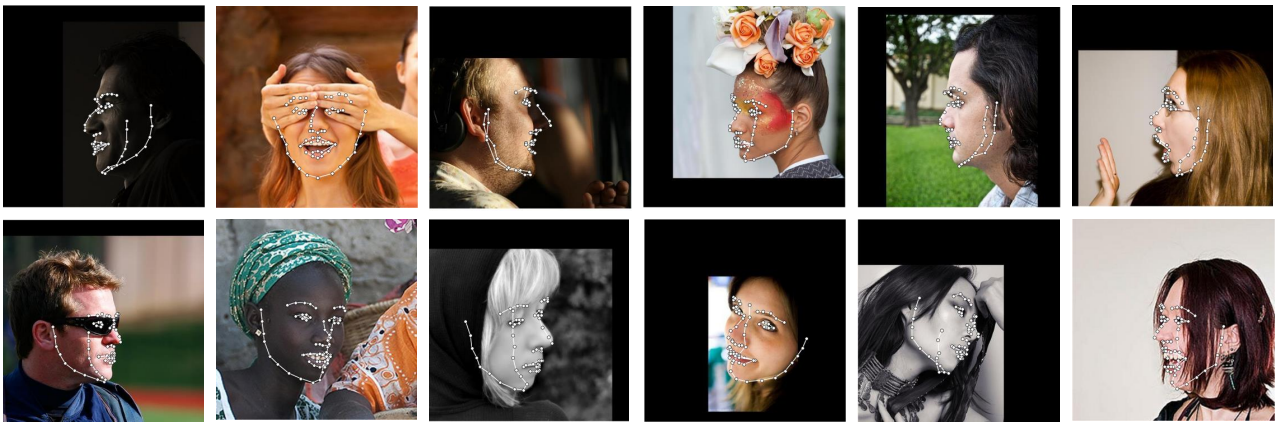
Results. Table 6 shows the obtained results. Similarly to our human pose estimation and face alignment experiments, we observe that the binarized network based on the proposed block significantly outperforms a similar-sized network constructed using the original bottleneck block, almost matching the performance of the real-valued network. Most of the performance improvement is due to the higher representation/learning capacity of our block, which is particularly evident for difficult cases like unusual poses, occlusions or challenging lighting conditions. For visual comparison, see Fig. 4.

Network type	pixel acc.	mean acc.	mean IU
Real, bottleneck	97.98%	77.23%	69.29%
Binary, bottleneck	97.41%	70.35%	62.49%
Binary, Ours	97.91%	76.02%	68.05%

Table 6: Results on 300W (Indoor&Outdoor). The pixel acc., mean acc. and mean IU are computed as in [8].

4. Visual results

This section provides qualitative results for our human pose estimation, face alignment and facial part segmentation experiments.



(a) Fitting examples produced by our binary network on AFLW2000-3D dataset. Notice that our method copes well with extreme poses and facial expressions and lighting conditions.

(b) Examples of human poses obtained using our binary network. Observe that our method produces good results for a wide variety of poses and occlusions.

Figure 3: Qualitative results produced by our method on (a) AFLW2000-3D and (b) MPII datasets.

References

[1] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[2] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[4] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv*, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[6] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *ICCV*, 2015.

[7] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[9] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016.

[10] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[11] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

Figure 4: Qualitative results on 300W (Indoor&Outdoor). Observe that the proposed binarized network significantly outperforms the original binary one, almost matching the performance of the real-valued network.