Supplementary Material for Deep Imbalanced Attribute Classification using Visual Attention Aggregation

Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris

Computational Biomedicine Lab
University of Houston
{nsarafia, xxu18, ikakadia}@central.uh.edu

Training Details

Since in both datasets we used a pre-trained primary network we first froze its weights and learned the attention masks using their corresponding loss function. This was done, to avoid back-propagating large prediction errors from the attention masks to the pre-trained network. After a few epochs of training solely the attention mechanism, the primary network is then unfrozen and trained end-to-end to produce multi-attribute predictions. For the WIDER-Attribute dataset we set the learning rate equal to 0.001 and use SGD with momentum set to 0.9 and a weight decay equal to 0.0005. The learning rate was divided by 10 (until 0.00001) when the error plateaus in the validation set. During pre-processing, we resized all images to 256×256 and extracted random crops of $[128, 224]$ (along with random mirroring and data shuffling) which were then resized to 224×224 and provided as an input to the network. For the PETA dataset we used Adam since it consistently outperformed SGD with a starting learning rate equal to 0.0001 with the same weight decay but with larger crops (in the range $[160, 224]$). In both datasets, the batch size was set to 32. We used MXNet/Gluon as our deep learning framework and a single NVIDIA GeForce GTX 1080 Ti GPU.

Architecture Details

Our backbone architecture is a ResNet-101 that extracts feature representations of dimensionality $7 \times 7 \times 2048$ which are then fed to a fully-connected layer initialized with Xavier initialization. Its dimensionality is equal to the number of classes denoted by $\mathcal{C}_l$ which for the WIDER dataset is equal to 14. The attention modules are placed on "stage3_activation22" and "stage4_activation2". Let Ck denote a Convolution-BatchNorm-ReLU layer with k filters and kernel size equal to 1 and Dk a fully-connected layer with k neurons. The attention module consists of C256-C256 and a convolutional layer with $\mathcal{C}_l$ number of filters. Its output is first spatially normalized and then multiplied by the output of the confidence weighting layer which is simply a convolutional layer with $\mathcal{C}_l$ number of filters and a sigmoid activation function. The output of the attention modules is fed to a C256-C512-C512-D$\mathcal{C}_l$ subnetwork the last convolutional layer of which has a kernel size equal to the spatial dimensions.