Large-Margin Softmax Loss for Convolutional Neural Networks

Weiyang Liu\textsuperscript{1*}, Yandong Wen\textsuperscript{2*}, Zhiding Yu\textsuperscript{3} and Meng Yang\textsuperscript{4}

\textsuperscript{1}Peking University \quad \textsuperscript{2}South China University of Technology \quad \textsuperscript{3}Carnegie Mellon University \quad \textsuperscript{4}Shenzhen University \quad \textsuperscript{*equal contribution}

Introduction

Cross-entropy loss together with softmax is arguably one of the most common used supervision components in convolutional neural networks (CNNs). Despite its simplicity, popularity and excellent performance, the component does not explicitly encourage discriminative learning of features. In this paper, we propose a generalized large-margin softmax ($L$-Softmax) loss which explicitly encourages intra-class compactness and inter-class separability between learned features. Moreover, $L$-Softmax not only can adjust the desired margin but also can avoid overfitting. We also show that the L-Softmax loss can be optimized by typical stochastic gradient descent. Extensive experiments on four benchmark datasets demonstrate that the deeply-learned features with L-Softmax loss become more discriminative, hence significantly boosting the performance on a variety of visual classification and verification tasks.

From Softmax Loss to Large-Margin Softmax Loss

Standard softmax loss can be written as

$$L = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{e^{y_i x_i}}{\sum_{j} e^{y_j x_i}} \right)$$

Using the transformation of inner products, the softmax loss is reformulated as

$$L_{i} = - \log \left( \frac{e^{\langle W_{i} \rangle_{n} \cdot x_{i} + b_{i}}}{\sum_{j} e^{\langle W_{j} \rangle_{n} \cdot x_{i} + b_{j}}} \right)$$

The large-margin softmax loss is formulated as

$$L_{i} = - \log \left( e^{\langle W_{i} \rangle_{n} \cdot x_{i} + b_{i}} + \sum_{j \neq i} e^{\langle W_{j} \rangle_{n} \cdot x_{i} + b_{j}} \right)$$

where $\psi(\theta) = (-1)^{k} \cos(m \theta) - \frac{2k}{m} \cos \left( \frac{(k+1) \theta}{m} \right)$$

Intuition & Geometric Interpretation

The purpose of L-Softmax loss is to learn discriminative features with large angular margin. We train the CNN with L-Softmax loss on MNIST dataset. The deeply-learned features are visualized in the following figure.

Experiments & Results

We perform extensive experiments on visual classification and face verification task, achieving state-of-the-art results on MNIST, CIFAR10, CIFAR100 and LFW public datasets.

- On all these datasets, we have shown that the classification accuracy will be improved with larger $m$, namely when the desired decision margin is set to be larger.
- The confusion matrix on CIFAR10, CIFAR10+ and CIFAR100 validate the discriminativeness of the deeply-learned features via our proposed L-Softmax loss.

\begin{tabular}{|c|c|}
\hline
Method & Error Rate \tabularnewline
\hline
CNN (Jain et al., 2009) & 0.53 \tabularnewline
DropConnect (Wan et al., 2013) & 0.57 \tabularnewline
FiNet (Romero et al., 2015) & 0.51 \tabularnewline
NIN (Lin et al., 2014) & 0.47 \tabularnewline
Maxout (Goodfellow et al., 2013) & 0.45 \tabularnewline
DSN (Lee et al., 2015) & 0.39 \tabularnewline
R-CNN (Li & Hu, 2015) & 0.31 \tabularnewline
GenPool (Lee et al., 2016) & 0.21 \tabularnewline
\hline
Hinge Loss & 0.47 \tabularnewline
Softmax & 0.40 \tabularnewline
L-Softmax (m=2) & 0.32 \tabularnewline
L-Softmax (m=3) & 0.31 \tabularnewline
L-Softmax (m=4) & 0.31 \tabularnewline
\hline
\end{tabular}

Accuracy (%) on MNIST

\begin{tabular}{|c|c|c|}
\hline
Method & Error Rate \tabularnewline
\hline
CNN (Jain et al., 2009) & 0.53 \tabularnewline
DropConnect (Wan et al., 2013) & 0.57 \tabularnewline
FiNet (Romero et al., 2015) & 0.51 \tabularnewline
NIN (Lin et al., 2014) & 0.47 \tabularnewline
Maxout (Goodfellow et al., 2013) & 0.45 \tabularnewline
DSN (Lee et al., 2015) & 0.39 \tabularnewline
R-CNN (Li & Hu, 2015) & 0.31 \tabularnewline
GenPool (Lee et al., 2016) & 0.21 \tabularnewline
\hline
Hinge Loss & 0.47 \tabularnewline
Softmax & 0.40 \tabularnewline
L-Softmax (m=2) & 0.32 \tabularnewline
L-Softmax (m=3) & 0.31 \tabularnewline
L-Softmax (m=4) & 0.31 \tabularnewline
\hline
\end{tabular}

Accuracy (%) on CIFAR100

Confusion matrix on CIFAR10, CIFAR10+, CIFAR100

Verification accuracy (%) on LFW

Acknowledgments

This work was supported by the National Science Foundation of China (No. 61872360, 61432019, 61473268, 61521064, 61272177) and the Fundamental Research Funds for the Central Universities (No. 2016KFXT14).