Rethinking Feature Distribution for Loss Functions in Image Classification

Weitao Wan, Yuanyi Zhong, Tianpeng Li, Jiansheng Chen
Department of Electronic Engineering, Tsinghua University

(A) Contributions

- We propose a large-margin Gaussian Mixture (L-GM) loss for image classification tasks, which is established on the assumption that the deep features of the training set follow a Gaussian Mixture distribution.
- We can easily introduce a classification margin into the L-GM loss and model the training feature distribution.
- Besides classification, it can be readily used to distinguish abnormal inputs, such as the adversarial examples, based on their features' likelihood to the training feature distribution.

(B) GM loss formulation

(1) The classification loss

The deep feature x follows a Gaussian mixture distribution.

$$p(x) = \sum_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k) p(k)$$

The conditional probability distribution and posterior probability distribution can be expressed by

$$p(x_i|z_i) = \mathcal{N}(\mathrm{x_i}; \mu_{z_i}, \Sigma_{z_i})$$

$$p(z_i|x_i) = \frac{\mathcal{N}(\mathrm{x_i}; \mu_{z_i}, \Sigma_{z_i}) p(z_i)}{\sum_{k=1}^{K} \mathcal{N}(x_i; \mu_k, \Sigma_k) p(k)}$$

As such, **the classification loss is**

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\mathcal{N}(\mathrm{x_i}; \mu_{z_i}, \Sigma_{z_i}) p(z_i)}{\sum_{k=1}^{K} \mathcal{N}(x_i; \mu_k, \Sigma_k) p(k)}$$

(2) The likelihood regularization

The log likelihood for the complete dataset $\{X, Z\}$ is

$$\log p(X, Z|\mu, \Sigma) = -\sum_{i=1}^{N} \left(\log \mathcal{N}(\mathrm{x_i}; \mu_{z_i}, \Sigma_{z_i}) + \log p(z_i) \right)$$

We define the likelihood regularization term as

$$\mathcal{L}_{lkd} = -\sum_{i=1}^{N} \log \mathcal{N}(\mathrm{x_i}; \mu_{z_i}, \Sigma_{z_i})$$

Finally the proposed GM loss is

$$\mathcal{L}_{GM} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{lkd}$$

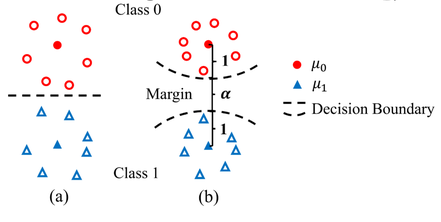
(C) Classification margin

Denote x_i's contribution to the classification loss as $\mathcal{L}_{cls,i}$.

$$\mathcal{L}_{cls,i} = -\log \frac{p(z_i)|\Sigma_{z_i}|^{-\frac{1}{2}} e^{-d_{z_i}}}{\sum_k p(k)|\Sigma_k|^{-\frac{1}{2}} e^{-d_k}}$$

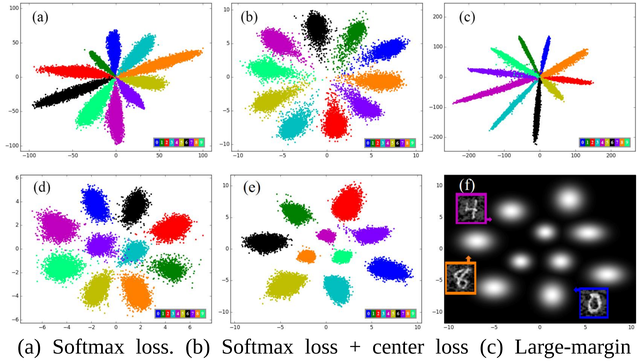
in which $$d_k = (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)/2$$

To introduce the classification margin, **we only need to substitute $d_{z_i} + m$ for d_{z_i}, in which m is the margin.** And we let $m = \alpha d_{z_i}$.



(D) Feature visualization

We plot the learned 2-d features of MNIST training set with LeNet for different loss functions.



(a) Softmax loss. (b) Softmax loss + center loss (c) Large-margin softmax loss. **(d) GM Loss without margin ($\alpha = 0$). (e) Large-margin GM loss ($\alpha = 1$).** (f) Heatmap of the learned likelihood.

(E) Image classification experiments

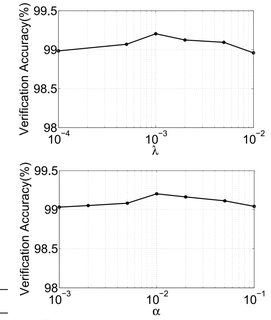
Extensive experiments on various recognition benchmarks like MNIST, CIFAR, ImageNet and LFW demonstrate the effectiveness of our proposal.

Method	Training Data	Accuracy
FaceNet [26]	200M	**99.65**
Deepid2+ [29]	0.3M	98.70
Softmax	0.49M	98.56 ± 0.03
L-Softmax [22]	0.49M	98.92 ± 0.03
Center [32]	0.49M	99.05 ± 0.02
LGM ($\alpha = 0.001$)	0.49M	99.03 ± 0.03
LGM ($\alpha = 0.005$)	0.49M	99.08 ± 0.02
LGM ($\alpha = 0.01$)	0.49M	$\mathbf{99.20 \pm 0.03}$

Table 5. Face verification performances on LFW of a single model. The 6 models at bottom are trained on our scheme while the 2 results on top are reported from the original paper.

Loss	1-crop		10-crop	
	top-1	top-5	top-1	top-5
Softmax	23.5±0.2	7.55±0.08	22.6±0.2	6.92±0.04
L-GM	**22.7±0.2**	**7.14±0.08**	**21.9±0.1**	**6.05±0.03**

Table 4. Error rates (%) on ILSVRC2012 validation set. For L-GM, we set α=0.01 and λ=0.1.



Parameter sensitivity on LFW. With a fixed $\alpha = 0.01$ (top) or a fixed $\lambda = 0.1$ (bottom)

(F) Adversarial verification experiments

The adversarial examples have low likelihood for the learned feature distribution. We can distinguish them from normal inputs by the likelihood.

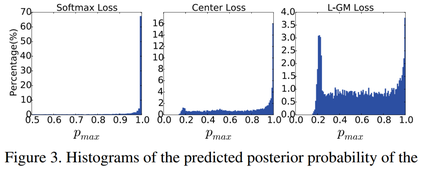


Figure 3. Histograms of the predicted posterior probability of the adversarial examples.

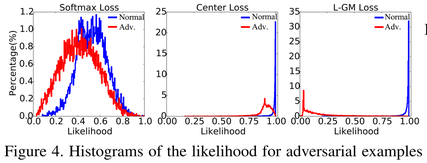


Figure 4. Histograms of the likelihood for adversarial examples (Adv.) and normal inputs (Normal).

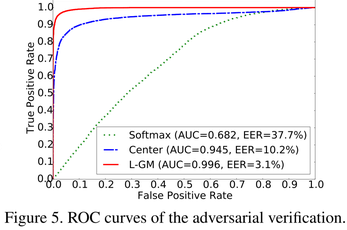


Figure 5. ROC curves of the adversarial verification.

GitHub