Rethinking Feature Distribution for Loss Functions in Image Classification

Weitao Wan et al., CVPR 2018

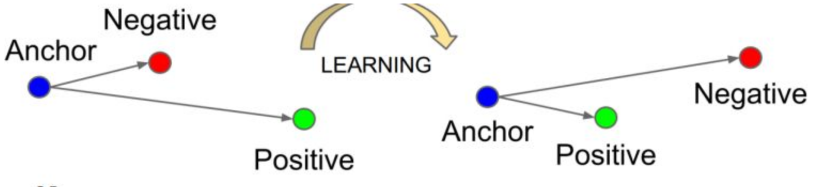
2019/12/23 Kangyeol Kim

Previous methods to improve the softmax loss

- Contrastive loss [1]

$$\mathrm{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

- Triplet loss [2]



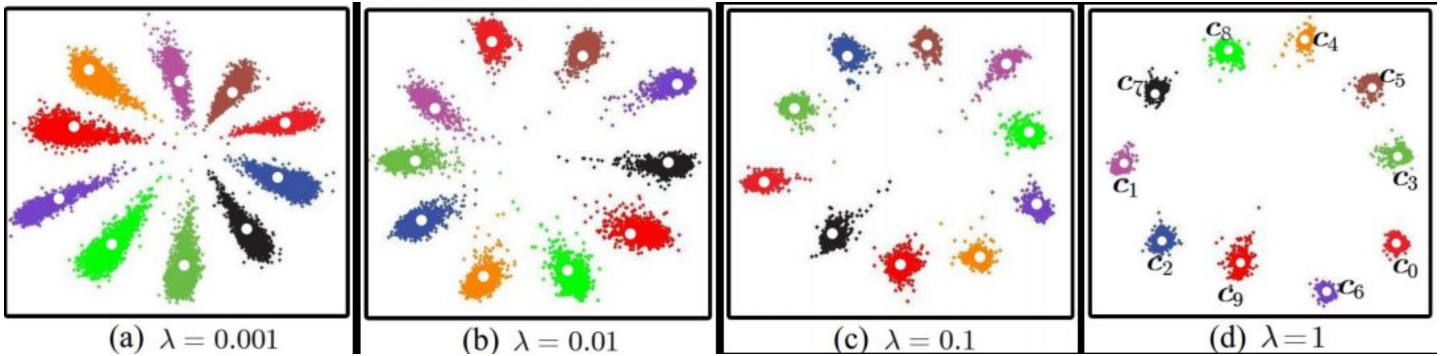
$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

- Problem: Explosion in the number of image pairs

[1] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification. NIPS 2014
[2] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. CVPR 2015.

Previous methods to improve the softmax loss

- Center loss [3]



$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

- Problem: scale problem

[1] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition. ECCV 2016

Previous methods to improve the softmax loss

- Large-margin softmax loss [4] - Positive “m” - Add Classification Margin!

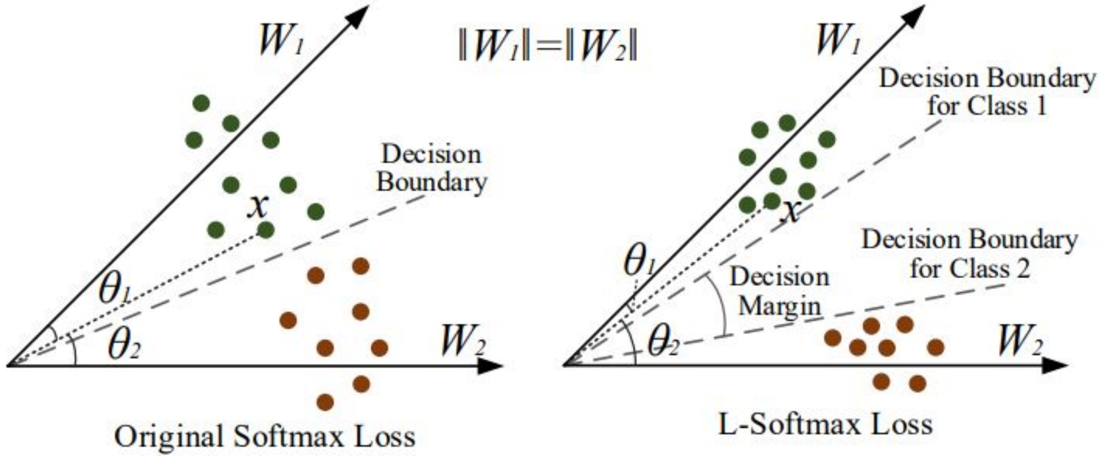
$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta}}\right. \quad \psi(\theta) = \begin{cases} \cos(m\theta), & 0 \leq \theta \leq \frac{\pi}{m} \\ \mathcal{D}(\theta), & \frac{\pi}{m} < \theta \leq \pi \end{cases}$$

[4] Liu W, Wen Y, Yu Z, et al. Large-Margin Softmax Loss for Convolutional Neural Networks. ICML 2016.

Previous methods to improve the softmax loss

- Large-margin softmax loss [4] - Geometric interpretation



[4] Liu W, Wen Y, Yu Z, et al. Large-Margin Softmax Loss for Convolutional Neural Networks. ICML 2016.

Summaries of the paper

- **Existing losses fail to generate likelihood in terms of probabilistic viewpoint**

- **GM loss improves the generalization capability of the trained model**

- **GM loss outputs calibrated scores.**

Methods - Gaussian Mixture Loss

- Under assumption that features follow gaussian mixture distribution

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

- Then,

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

-
- A Classification loss can be computed as the cross-entropy between the posterior probability distribution and the one-hot class label

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

Methods - Large Margin GM Loss

- Denote x(i)'s contribution to the classification loss is:

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

- Introducing a margin term to this loss,

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

- Intuition: assume / p(k), Sigma(k) are identical for all the classes, x(i) should be closer to the feature mean of class z(i) than to that of the other classes by at least "m"

$$e^{-d_{z_i} - m} > e^{-d_k} \iff d_k - d_{z_i} > m \quad , \forall k \neq z_i$$

Methods - Geometric interpretation

- Adapting $m = \alpha d_{z_i}$

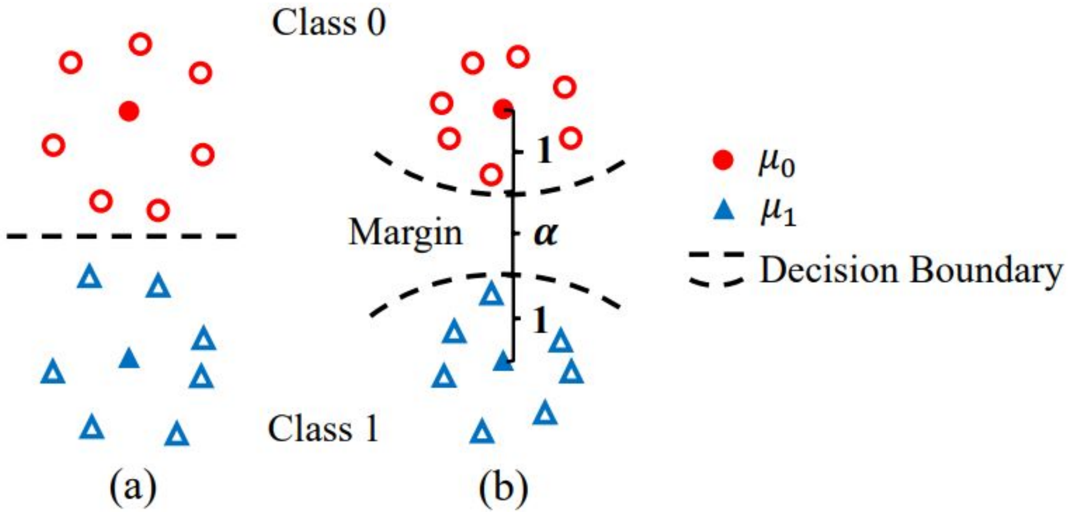


Figure 2. A geometry interpretation of the relationship between α and the margin size in the training feature space using (a) GM loss without margin $\alpha = 0$; (b) large-margin GM loss with $\alpha > 0$.

Methods - Forcing feature to follow GMM

- Adapting the likelihood regularization term:

$$p(\mathrm{X}, \mathrm{Z}|\mu, \Sigma) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathbb{1}(z_i = k)\mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i})p(z_i)$$

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

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

- Comparison with center loss: center loss is a special case of likelihood reg.
- More accurate likelihood estimation

Experiments - Qualitative result

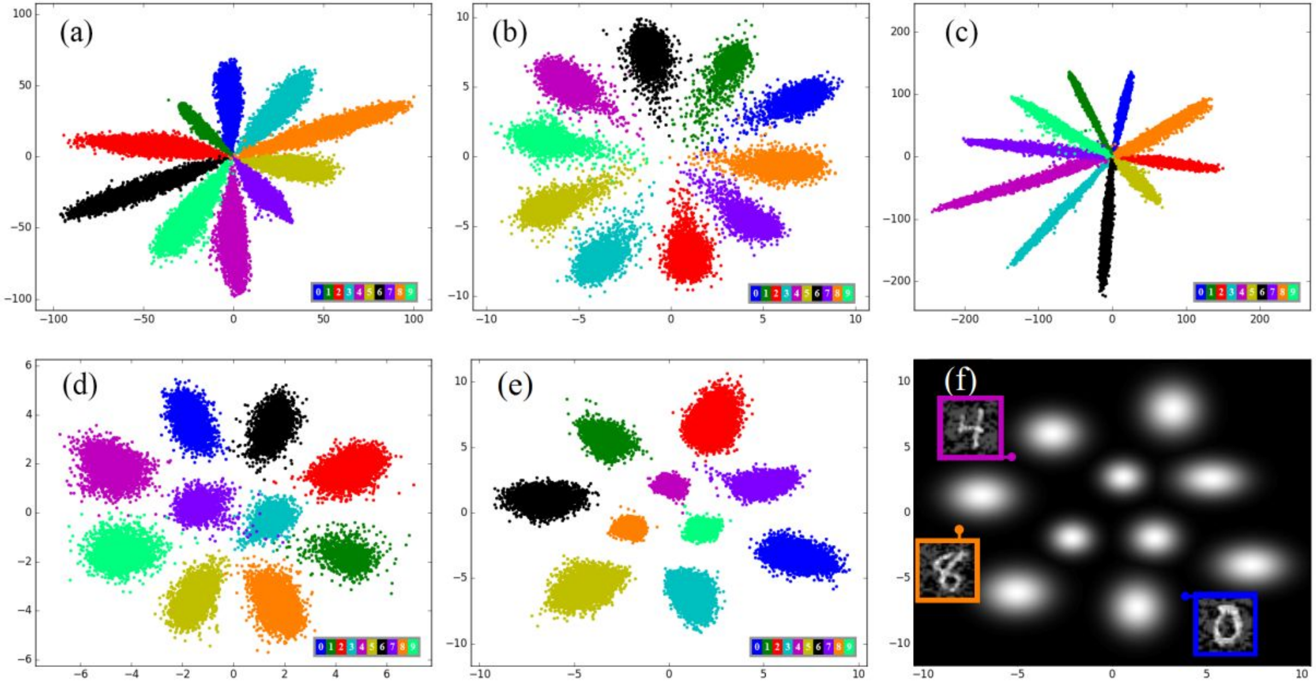


Figure 1. Two-dimensional feature embeddings on MNIST training set. (a) Softmax loss. (b) Softmax loss + center loss [32]. (c) Large-margin softmax loss [22]. (d) GM Loss without margin ($\alpha = 0$). (e) Large-margin GM loss ($\alpha = 1$). (f) Heatmap of the learned likelihood corresponding to (e). Higher values are brighter. Several adversarial examples generated by the Fast Gradient Sign Method [8] have extremely low likelihood according to the learned GM distribution and thus can be easily distinguished. This figure is best viewed in color.

Experiments - Quantitative result (1/3) - Accuracy

Loss Functions	C100	C100+
Center [32]	24.85 ± 0.06	21.05 ± 0.03
L-Softmax [22]	24.83 ± 0.05	20.98 ± 0.04
Softmax	25.61 ± 0.07	21.60 ± 0.04
LGM($\alpha = 0.1$)	23.74 ± 0.08	20.94 ± 0.03
LGM($\alpha = 0.2$)	$\mathbf{23.04 \pm 0.08}$	20.85 ± 0.04
LGM($\alpha = 0.3$)	23.80 ± 0.06	$\mathbf{20.76 \pm 0.03}$

Table 3. Recognition error rates (%) on CIFAR-100 using a VGG-like 13 layer CNN with different loss functions.

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.

Experiments - Quantitative result (1/3) - Adversarial Examples

ϵ	Softmax	Center	L-GM($\alpha = 1$)
0	0.68	0.47	0.39
0.1	24.08	43.13	23.63
0.2	75.56	67.17	64.40
0.3	84.87	85.49	81.62

Table 6. Classification error rates (%) on adversarial examples generated from the MNIST test set using FGSM. $\epsilon = 0$ means that the inputs are normal MNIST test images.

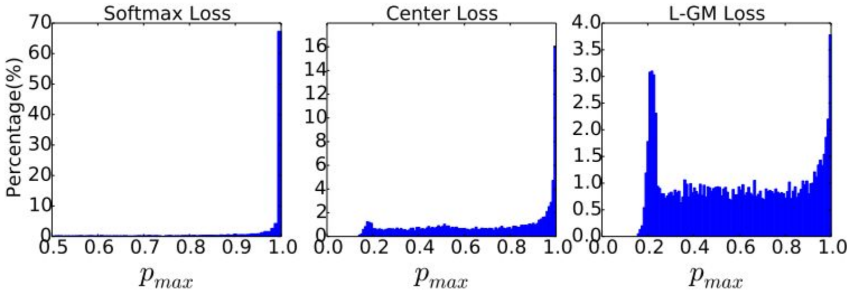


Figure 3. Histograms of the predicted posterior probability of the adversarial examples.

Experiments - Quantitative result (1/3) - Adversarial verification

- Likelihood of each loss
  - Softmax loss = $l_{S,i} = w_{\hat{z}_i}^T x_i + b_{\hat{z}_i}$
  - Center loss = $l_{C,i} = exp(-\|x_i - \mu_{\hat{z}_i}\|^2/2)$
  - GM loss = $l_{GM,i} = exp(-\|x_i - \mu_{\hat{z}_i}\|^2/2)$

Lemma 1. *If $\Sigma_k = I$ (identity matrix), $p(k) = 1/K, \forall k \in [1, K]$, the center loss $\mathcal{L}_C$ and the likelihood regularization $\mathcal{L}_{lkd}$ satisfy Eq. 16, in which D is the feature dimension.*

$$\mathcal{L}_{lkd} = \mathcal{L}_C + \frac{N}{2} D \log(2\pi) \tag{16}$$

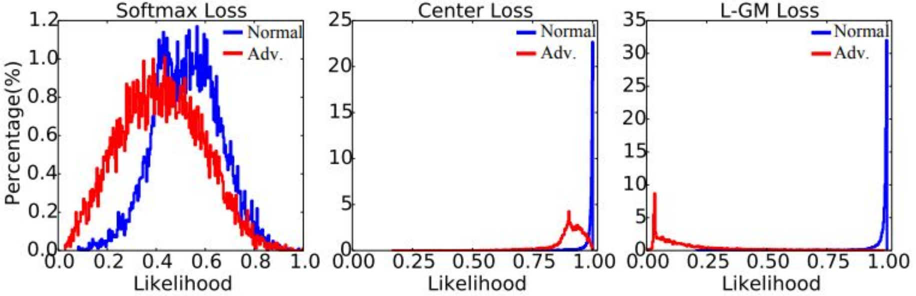


Figure 4. Histograms of the likelihood for adversarial examples (Adv.) and normal inputs (Normal).

THANK YOU!