Squared Earth Mover's Distance Loss for Training Deep Neural Networks on Ordered-Classes

Le Hou
Dept. of Computer Science
Stony Brook University

Chen-Ping Yu
Phiar Technologies, Inc

Dimitris Samaras
Dept. of Computer Science
Stony Brook University

Abstract

In the context of multi-class single-label classification, the loss function of deep learning methods compares the predicted class distribution versus the ground truth class distribution. The commonly used cross-entropy loss ignores the intricate inter-class relationships that often exist in real-life tasks such as age classification. We propose to leverage these relationships between classes by training deep nets with the exact squared Earth Mover's Distance (also known as Wasserstein distance), assuming that the classes are ordered: one can put all classes in a one-dimensional space such that the dissimilarities between classes are represented by the euclidean distances between them. The EMD^2 loss uses the predicted probabilities of all classes and penalizes the miss-predictions according to the dissimilarities between classes. Our exact EMD^2 loss yields state-of-the-art results with limited computational overhead on age estimation and image aesthetics datasets.

1 Introduction

Deep neural networks (DNNs) have become the preferred method for most machine learning applications [14, 33, 34, 24, 13, 5]. In general, most DNNs are trained under one of two tasks: regression and classification. In a regression task, the network learns to generate a real-valued output that matches the ground-truth [2, 6]. In a classification task, the network learns to categorize an input to one of the training classes [8, 37, 14, 33, 23].

To train a multi-class single-label classification network, softmax cross-entropy loss is by far the most popular loss function for the training regime, where the ground-truth is a binary vector consisting of a value 1 at the correct class index, and 0s everywhere else [20, 19]. During training, the objective is to minimize the negative log-likelihood of the loss by multiplying the network's predictions to the binary ground-truth vectors. This loss function does not take into account inter-class relationships which can be very informative. For example, we want to estimate age-groups from face images. In Fig. 1, two predicted class distributions have identical softmax cross-entropy loss. However, one is clearly more preferable than the other.

In this work, we show how the exact squared Earth Mover's Distance (EMD) [31] can be applied as a stand-alone loss function for multi-class single-label classification problems using CNNs. The EMD is also known as the Wasserstein distance [1, 3], which is the minimal cost required to transform one distribution to another [31]. Recent work formulated an approximate Wasserstein loss for supervised multi-class multi-label learning [10, 28]. In contrast, we show that an *exact* (without approximation) squared EMD (EMD^2) loss exists for training single-label deep learning models directly, assuming that the classes are ordered: one can put all classes in a one-dimensional space such that the dissimilarities between classes are represented by the euclidean distances between them. We choose to use EMD^2 instead of EMD as the loss function for faster convergence with gradient descent [32, 25]. Our experiments show that CNNs trained with our EMD^2 loss perform better than CNNs with the standard softmax cross-entropy loss. We verify our approach on datasets with known

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

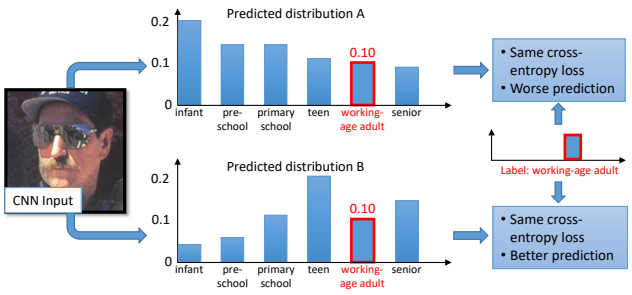

Figure 1: In many classification tasks, there are relationships or even orderings between classes. However the cross-entropy loss ignores these relationships and only focuses on the predicted probability of the ground truth class. In this example, the two given predicted distributions have the same cross-entropy loss. But clearly predicted distribution B is preferable to A.

ordered-classes [8, 9, 11, 30, 7, 18, 35, 26]. For the first time, we show how an *exact* EMD^2 loss function can be used to train CNNs on ordered-classes classification problems. Our method achieves state-of-the-art results on the Adience dataset [8], the Image of Groups dataset [11], and the image aesthetics with attributes database (AADB) [18] without using additional image attributes such as color harmony.

2 EMD^2 loss on ordered-classes

We first introduce the softmax cross-entropy loss and some other notation. For a single-label classification problem with C classes, a network's softmax layer outputs a probability distribution $\mathbf{p}$ of length C, with its i-th entry $\mathbf{p}_i$ being the predicted probability of the i-th class. The soft-max guarantees that $\sum_i \mathbf{p}_i = 1$. We denote the ground truth as a binary vector $\mathbf{t}$ of length C. Also $\sum_i \mathbf{t}_i = 1$. Given a training example, the cross-entropy loss between the prediction $\mathbf{p}$ and the ground truth vector $\mathbf{t}$ is defined as $\mathrm{E}_\mathrm{X}(\mathbf{p}, \mathbf{t}) = -\sum_{i=1}^{C} \big(\mathbf{t}_i \log(\mathbf{p}_i)\big)$. We assume that the k-th class is the ground truth label: $\mathbf{t}_k = 1$ and $\mathbf{t}_i = 0$ for $i \neq k$. Thus the differentiation of $\mathrm{E}_\mathrm{X}(\mathbf{p}, \mathbf{t})$ is: $\mathrm{E}'_X(\mathbf{p}, \mathbf{t}) = -\mathbf{p}'_k / \mathbf{p}_k$. The backpropagation of a DNN with cross-entropy loss only depends on $\mathbf{p}_k$. This is less robust compared to a loss function that depends on all entries of $\mathbf{p}$ as argued in Fig. 1.

We assume that a CNN should ideally predict class distributions such that classes closer to the ground truth class should have higher predicted probabilities than classes that are further away. We formulate this using the Earth Mover's Distance (EMD). The EMD is defined as the minimum cost to transport the mass of one distribution (histogram) to the other.

Mass transportation defines the problem of transporting mass from a set of supplier clusters to a set of consumer clusters. Its formal definition [31] is: Let $\mathbf{p} = \{(\mathbf{a}_1, \mathbf{p}_1), (\mathbf{a}_2, \mathbf{p}_2), \ldots, (\mathbf{a}_C, \mathbf{p}_C)\}$ be the supplier signature (distribution or histogram) with C clusters (bins), where $\mathbf{a}_i$ represents each cluster and $\mathbf{p}_i$ is the mass (value) in each cluster. Let $\mathbf{t} = \{(\mathbf{b}_1, \mathbf{t}_1), (\mathbf{b}_2, \mathbf{t}_2), \ldots, (\mathbf{b}_{C'}, \mathbf{t}_{C'})\}$ be the consumer signature. Let $\mathbf{D}$ be the ground distance matrix where its i, j-th entry $\mathbf{D}_{i,j}$ is the distance between $\mathbf{a}_i$ and $\mathbf{b}_j$. Matrix $\mathbf{D}$ is usually defined as the l-norm distance between clusters: $\mathbf{D}_{i,j} = ||\mathbf{a}_i - \mathbf{b}_j||_l$. Let $\mathbf{F}$ be the transportation matrix where its i, j-th entry $\mathbf{F}_{i,j}$ indicates the mass transported from $\mathbf{a}_i$ to $\mathbf{b}_j$. A valid transportation satisfies four constraints. First, the amount of mass transported must be positive. Second, the amount of mass transported from a supplier cluster $\mathbf{p}_i$ must not exceed its total mass. Third, the amount of mass transported to a consumer cluster $\mathbf{t}_j$ must not exceed its total mass. Finally, the total flow must not exceed the total mass that can be transported. These four conditions can be summarized respectively below:

$$\text{for all } i,j,\ \mathbf{F}_{i,j} \geq 0;\ \sum_{j=1}^{C'} \mathbf{F}_{i,j} \leq \mathbf{p}_i;\ \sum_{i=1}^{C} \mathbf{F}_{i,j} \leq \mathbf{t}_j;\ \sum_{i=1}^{C} \sum_{j=1}^{C'} \mathbf{F}_{i,j} = \min\Big(\sum_{i=1}^{C} \mathbf{p}_i, \sum_{j=1}^{C'} \mathbf{t}_j\Big). \quad (1)$$

Under the constraints defined above, the overall cost of flow $\mathbf{F}$ is defined as: $\mathrm{W}(\mathbf{b}, \mathbf{t}, \mathbf{F}) = \sum_i \sum_j \mathbf{D}_{i,j} \mathbf{F}_{i,j}$. The EMD between two vectors, denoted as $\mathrm{EMD}(\mathbf{p}, \mathbf{t})$ is the minimum cost of work that satisfies the constraints, normalized by the total flow: $\mathrm{EMD}(\mathbf{p}, \mathbf{t}) = \inf_{\mathbf{F}} \dfrac{\mathrm{W}(\mathbf{b}, \mathbf{t}, \mathbf{F})}{\sum_i \sum_j \mathbf{F}_{i,j}}$.

2.1 Ground distance matrix of ordered-classes

Computing the EMD between two distributions requires a predefined matrix, the ground distance matrix $\mathbf{D}$ which is unknown in most cases. However, in classification tasks with ordered classes we can define $\mathbf{D}$. The difference between ordered-class classification and regression is that in the problem of ordered-class classification, the ground truth labels and predictions are discrete. Often, a multi-class classification model [12] performs better than a regression model.

Without loss of generality, we assume that in all ordered-class classification problems, the classes are ranked as $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \ldots, \mathbf{t}_C$ and the distance between $\mathbf{t}_i$ and $\mathbf{t}_j$ is $|i-j|$.

2.2 EMD^2 loss for ordered-class classification

EMD has been shown to be equivalent to Mallows distance which has a closed-form solution [22], if the ground distance matrix $\mathbf{D}$ and distributions $\mathbf{p}$ and $\mathbf{t}$ satisfy certain conditions, as shown in [22]. We will show that these required conditions are satisfied in ordered-class classification problems.

The first condition is that the two distributions $\mathbf{p}$ and $\mathbf{t}$ must have equal mass: $\sum_i \mathbf{p}_i = \sum_j \mathbf{t}_j$. This condition is always satisfied and $C = C'$ if $\mathbf{p}$ is produced by a softmax layer. The second condition is that the ground distance matrix $\mathbf{D}$ must have an one-dimensional embedding. This assumption is always satisfied in ordered-class classification problems. The third and final condition is that the distributions to be compared must be sorted vectors. This condition is also always satisfied since we assumed $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_C$ and $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{C'}$ are sorted without loss of generality. Then, based on the conclusion by Levina et al. [22], the normalized EMD can be computed exactly and in closed-form: $\mathrm{EMD}(\mathbf{p}, \mathbf{t}) = \left(\frac{1}{C}\right)^{\frac{1}{l}} ||\mathrm{CDF}(\mathbf{p}) - \mathrm{CDF}(\mathbf{t})||_l$, where $\mathrm{CDF}(\cdot)$ is a function that returns the cumulative density function of its input.

We use $l = 2$ for Euclidean distance and also for $\mathbf{D}$. Dropping the normalization term, we obtain the final EMD^2 loss E_E as: $\mathrm{E}_E(\mathbf{p}, \mathbf{t}) = \sum_{i=1}^{C} \Big(\mathrm{CDF}_i(\mathbf{p}) - \mathrm{CDF}_i(\mathbf{t})\Big)^2$, where $\mathrm{CDF}_i(\mathbf{p})$ is the i-th element of the CDF of $\mathbf{p}$. This equation is directly applicable to ordered-class classification problems on neural networks trained with backpropagation.

3 Experiments and Results

We test the EMD^2 loss on different network architectures including AlexNet [19], the VGG 16-layer network [33], and the wide residual network [37]. For optimization, we use stochastic gradient descent with momentum 0.98 in all experiments. The learning rates were selected from $\{10^{-1.5}, 10^{-2}, 10^{-2.5}, 10^{-3}, 10^{-3.5}, 10^{-4}, 10^{-4.5}\}$ individually for each method on each dataset. For experiments on all datasets, during training we randomly crop, flip, rotate, adjust the RGB colors and aspect ratio of input images for data augmentation. During testing, we use the average prediction from the center crop and its mirrored image. Our implementation of EMD^2 loss functions increase the CNN training time less than 10% for each iteration and has the same test time. We use Theano [36] for network implementation. Code is available at `http://www3.cs.stonybrook.edu/~cvl/emd2.html`. More experimental results can be found at [16]

Methods tested: For network architectures, we train AlexNet (**ALX**) [19] from scratch, to compare with published baselines [21, 18]. For the Adience dataset, we test a smaller version of AlexNet following [21], which we call it **ALXs**. We also fine-tune the VGG 16-layer network (**VGG$_\mathbf{F}$**) [33] pre-trained on ImageNet, to compare with the published baseline [30]. Additionally, we train a 40-layer residual network (**RES**) with identity mapping and bottleneck design [37] from scratch. We also fine-tune a RES pre-trained on ImageNet and name it **RES$_\mathbf{F}$**.

For the loss functions, we test the softmax cross-entropy loss (**XE**), the $L2$ regression loss (**REG**), the approximated EMD loss (**AEMD**) and the proposed EMD^2 loss (**EMD**). For the REG loss, the output neuron of a regression network uses a linear activation function, following the conventional regression CNN approach [2]. For the AEMD loss, we use the euclidean distances between class centers (represented as CNN features) as the required ground distance matrix $\mathbf{D}$. The number of

	ALXs	VGG_F	RES	RES_F		Others
	AEM/AEO	AEM/AEO	AEM/AEO	AEM/AEO		AEM/AEO
XE	53.0/85.7	60.9/92.8	58.1/90.3	60.1/92.1	ALXs by[21]	50.7/84.7
REG	49.8/85.1	56.5/94.0	57.3/91.8	60.8/**94.3**	CascadeNN[4]	52.9/88.5
EMD	57.0/90.4	59.2/92.6	61.9/93.1	**62.2/94.3**	VGG_FDEX[30]	55.6/89.7
AEMD	53.9/88.7	60.1/93.4	58.7/91.7	59.6/92.5	SAAF[15]	**61.3/95.1**

Table 1: Accuracy of exact match (AEM%) and with-in-one-category-off match (AEO%) results on the Adience dataset [8]. Note that better results are reported [30, 15] using an external age estimation dataset IMDB-WIKI for training, which is 10 times larger than Adience.

	VGG_F	RES	RES_F		Others
	AEM/AEO	AEM/AEO	AEM/AEO		AEM/AEO
XE	64.3/95.6	60.0/91.5	61.8/94.2	Multi-CNN [7]	56/92
REG	60.2/**96.6**	52.8/92.2	61.4/95.7	SAAF [15]	54.2/93.0
EMD	**65.0**/96.1	59.3/92.5	63.1/95.3	Deep Attention [29]	**60.0/94.5**

Table 2: Accuracy of exact match (AEM%) and with-in-one-category-off match (AEO%) results on the Images of Groups dataset [11]. Our EMD^2 loss outperforms the cross-entropy loss and the $L2$ loss (regression) based methods in most cases, and improves the state-of-the-art.

	VGG_F	RES	RES_F	$VGG_F \times 8$		Others
XE	0.6283	0.5003	0.6693	-	ALX by [18]	0.5923
REG	0.6096	0.5235	0.6609	-	*Best of [18]	0.6782
EMD	0.6682	0.5448	**0.6768**	**0.6889**	*Aesthetic Network [27]	**0.6890**

Table 3: Spearmans' ρ results on the image aesthetics with attributes database (AADB) [18]. *: used additional 11 labels of image attributes such as color harmony, and image content information. Our EMD^2 loss outperforms cross-entropy loss and $L2$ loss based methods significantly. Averaging the results of eight VGG_F networks, we achieve a state-of-the-art result without image attributes.

matrix scaling iterations in [10] is set to 100. The entropic regularizer in [10] is selected from $\{0.1, 1, 10\}$ based on validation error. We use a Caffe implementation of this loss function [17].

Age estimation We test our method on the Adience age estimation dataset [8] which contains 26,000 images in 8 age-groups, and a five-fold cross-validation evaluation scheme. We compare with existing methods [21, 30, 15] that used ALXs and VGG_F. We evaluate using the conventional accuracy of exact match (AEM%) and with-in-one-category-off match (AEO%). The results are shown in Tab. 1. Our method improves the state-of-the-art when training without external face datasets. Because the $L2$ regression loss is sensitive to outliers, it achieves low AEM scores.

We also test our method on the Images of Groups dataset [11] which contains 3,500 training face images and 1,000 testing face images in 7 age-groups. In Tab. 2, our EMD^2 loss outperforms the cross-entropy loss and the $L2$ loss (regression) in most of the cases, improving the state-of-the-art.

Image aesthetics We test our method on the Image Aesthetics with Attributes Database (AADB) [18] which contains 8,458 training and 1,000 testing images, labeled as real numbers in $[0.0, 1.0]$. To transform this dataset into a classification dataset, we discretize the real number labels to 10 bins, balancing the number of training images in each bin. During testing, we compute the expected aesthetic scores according to the predicted distributions. This give us real-numbered predictions. We use Spearmans' rank correlation ρ as the evaluation metric, following [18].

The results are shown in Tab. 3. Our EMD^2 loss again outperforms cross-entropy loss and $L2$ loss (regression) significantly. We conduct additional experiments by discretizing the real-numbered aesthetic labels to 8 different number of bins (3,4,5,6,7,8,9,10 bins), which give us 8 sets of ground truth labels. Then, we fine-tune one VGG_F network with EMD loss for each ground truth set and average the prediction results into an ensemble model. It achieves state-of-the-art results training only on image data. The existing state-of-the-art method is trained using 11 labels such as color harmony and vivid color information, in addition to the image data.

Acknowledgments

This work is supported by a gift from Adobe and the Partner University Fund 4DVision project.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

[2] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. Robust optimization for deep regression. In *ICCV*, 2015.

[3] V. I. Bogachev and A. V. Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67, 2012.

[4] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa. A cascaded convolutional neural network for age estimation of unconstrained faces. In *Biometrics Theory, Applications and Systems (BTAS)*, 2016.

[5] L. Deng and J. C. Platt. Ensemble deep learning for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014.

[7] Y. Dong, Y. Liu, and S. Lian. Automatic age estimation based on deep learning algorithm. *Neurocomputing*, 187, 2016.

[8] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 2014.

[9] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV Workshop*, 2015.

[10] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. In *NIPS*, 2015.

[11] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009.

[12] P. Golik, P. Doetsch, and H. Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, 2013.

[13] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2015.

[15] L. Hou, D. Samaras, T. Kurc, Y. Gao, and J. Saltz. Convnets with smooth adaptive activation functions for regression. In *Artificial Intelligence and Statistics*, 2017.

[16] L. Hou, C.-P. Yu, and D. Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv*, 2016.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*, 2014.

[18] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. *arXiv preprint arXiv:1606.01621*, 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[21] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshop*, 2015.

[22] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, 2001.

[23] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[24] S. Liu, N. Yang, M. Li, and M. Zhou. A recursive recurrent neural network for statistical machine translation. In *ACL (1)*, 2014.

[25] D. G. Luenberger. *Introduction to linear and nonlinear programming*, volume 28. 1973.

[26] K. Ma, D. Samaras, M. Petrucci, D. L. Magnus, et al. Texture classification for rail surface condition evaluation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[27] G. Malu, R. S. Bapi, and B. Indurkhya. Learning photography aesthetics with deep cnns. *arXiv preprint arXiv:1707.03981*, 2017.

[28] M. Martinez, M. Haurilet, Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Relaxed earth mover's distances for chain-and tree-connected spaces and their use as a loss function in deep learning. *arXiv preprint arXiv:1611.07573*, 2016.

[29] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzàlez. Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 72, 2017.

[30] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2016.

[31] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40, 2000.

[32] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12, 2011.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014.

[34] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.

[35] D. Sutić, I. Brešković, R. Huić, and I. Jukić. Automatic evaluation of facial attractiveness. In *MIPRO, 2010 Proceedings of the 33rd International Convention*, 2010.

[36] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

[37] S. Zagoruyko and N. Komodakis. Wide residual networks. *BMVC*, 2016.