



AdaMS: Deep Metric Learning with Adaptive Margin and Adaptive Scale for Acoustic Word Discrimination

Myunghun Jung, Hoirin Kim

School of Electrical Engineering, KAIST, Daejeon, Republic of Korea

kss2517@kaist.ac.kr, hoirkim@kaist.ac.kr

Abstract

Many recent loss functions in deep metric learning are expressed with logarithmic and exponential forms, and they involve margin and scale as essential hyper-parameters. Since each data class has an intrinsic characteristic, several previous works have tried to learn embedding space close to the real distribution by introducing adaptive margins. However, there was no work on adaptive scales at all. We argue that both margin and scale should be adaptively adjustable during the training. In this paper, we propose a method called Adaptive Margin and Scale (AdaMS), where hyper-parameters of margin and scale are replaced with learnable parameters of adaptive margins and adaptive scales for each class. Our method is evaluated on Wall Street Journal dataset, and we achieve outperforming results for word discrimination tasks.

Index Terms: deep metric learning, adaptive margin, adaptive scale, acoustic word discrimination

1. Introduction

A successful approach to discriminating two spoken words based on phonetic similarity is to measure the distance between their neural representations, called *acoustic word embeddings* (AWEs) [1, 2]. Learning AWEs is subsumed into deep metric learning (DML), and the recent progress via introducing *acoustically grounded word embeddings* (AGWEs) [3, 4, 5, 6], which encode phonetic contents of text inputs, is also related to the great success of proxy-based DML [7, 8, 9]. Thus, finding the optimal DML or proxy-based DML method can lead to further improvement on acoustic word discrimination task.

In recent DML works that have focused on loss functions while leaving the backbone network the same, the loss functions are generally expressed as the sum of two terms making anchor-positives closer and anchor-negatives farther. For each term, logarithmic and exponential functions such as softmax, softplus, and log-sum-exp have widely been utilized to handle complicated relations within a batch [7, 8, 9, 10, 11]. They involve margin and scale, where the margin determines boundaries on embedding space, and the scale controls the intensity of punishment for violations. These hyper-parameters are uniformly tuned under the assumption that all classes have identical shapes of distributions. However, the consequently learned embedding space cannot describe the real distribution perfectly since each class has an intrinsic characteristic.

To address the problem, several methods have proposed introducing an adaptive margin. In [12], a margin for Triplet loss varies with the average distance between positive samples in an anchor class. In [13], an auxiliary network generates margins for classification loss from pre-trained semantic embeddings of anchor and negative classes. Since these kinds of methods with

margins dependent on certain quantities are deficient in generalization capability, an adaptive margin is defined for each class as a learnable parameter that is jointly optimized with the network [14, 15, 16, 17]. In [15, 16, 17], regularization is also employed to induce larger margins as they are preferred for higher discriminability.

While research on the adaptive margin is active as such, an adaptive scale is not considered at all, even though the scale is also the essential component. Therefore, in this paper, we try to apply the adaptive margin and adaptive scale together. Specifically, we propose a new but straightforward method called Adaptive Margin and Scale (AdaMS) to give flexibility in training dynamics of DML loss functions by replacing hyper-parameters of margin and scale with learnable parameters of adaptive margins and adaptive scales, respectively. Then we provide a gradient-based analysis of how the margins and scales change adaptively in affecting the training process simultaneously, which is described with examples. In addition, following the argument of [18, 19] that DML loss functions are sensitive to hyper-parameters, we impose non-linear constraints on the intervals where the adaptive margins and adaptive scales can vary.

Our AdaMS method is applied to the current state-of-the-art proxy-based DML loss function, and then the learned embedding space is evaluated on Wall Street Journal (WSJ) dataset for word discrimination tasks. We demonstrate the effectiveness of the proposed approach by achieving meaningful results outperforming all other methods.

2. Proposed method

In this section, we first review Asymmetric-Proxy (AsyP) loss [9], which reported highly improved results, as the baseline to which AdaMS is applied. We then describe our AdaMS method and analyze the behaviors of adaptive margins and adaptive scales separately based on their gradients for optimization and an illustration.

2.1. Review of Asymmetric-Proxy loss

Let $\{(\mathbf{x}_i, \mathbf{t}_i, c_i) | i = 1, 2, \dots, N\}$ be a batch of N data tuples, where \mathbf{x}_i is the AWE of the i -th speech segment, \mathbf{t}_i is the AGWE of the text label, and c_i is the word class index. Following the general formulation for proxy-based DML in [9], loss functions are given as the sum of the anchor-positive term \mathcal{L}_i^P and anchor-negative term \mathcal{L}_i^N for the i -th anchor:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (\mathcal{L}_i^P + \mathcal{L}_i^N). \quad (1)$$

AsyP loss [9] uses two different functions for \mathcal{L}^P and \mathcal{L}^N

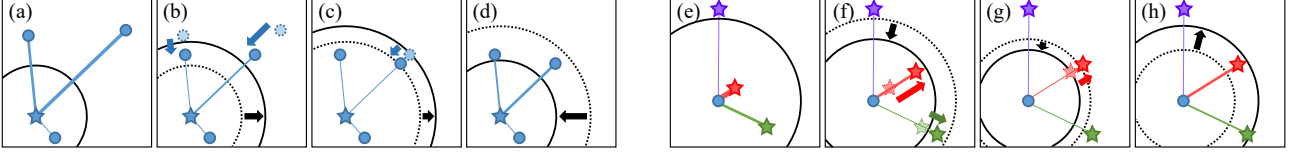


Figure 1: Illustration of adaptive margins for positives (a)-(d) and negatives (e)-(h). Small circles are AWEs, stars are AGWEs, and their colors represent distinct classes. The hardnesses are expressed by the thickness of edges. The black solid line indicates the boundary determined by the adaptive margin. All dashes show their previous states, and the movements are represented by arrows.

to show that the optimal function making anchor and positives closer has to be considered separately from one for anchor and negatives to be farther. Specifically, AsyP loss takes the positive term from Multi-Similarity (MS) loss [10] and the negative term from Binomial Deviance (BD) loss [11] with modification in computing similarities by introducing proxies, and then combines them. Each term is given as:

$$\mathcal{L}_i^{\mathcal{P}} = \frac{1}{\alpha} \log \left(1 + \sum_{j \in \mathcal{P}_i} e^{\alpha(\lambda - S(\mathbf{t}_i, \mathbf{x}_j))} \right), \quad (2)$$

$$\mathcal{L}_i^{\mathcal{N}} = \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \log \left(1 + e^{\beta(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda)} \right), \quad (3)$$

where $\mathcal{P}_i = \{j | c_j = c_i\}$, $\mathcal{N}_i = \{k | c_k \neq c_i\}$, and $S(\cdot, \cdot)$ denotes the cosine similarity. λ is a fixed margin, $\alpha > 0$ is a fixed scale for the anchor-positive term, and $\beta > 0$ is a fixed scale for the anchor-negative term. Notice that the proxies \mathbf{t} are utilized as anchors in $\mathcal{L}^{\mathcal{P}}$ and as negatives in $\mathcal{L}^{\mathcal{N}}$. The gradient of AsyP loss with respect to S is then:

$$\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial S(\mathbf{t}_i, \mathbf{x}_j)} = \frac{-e^{\alpha(\lambda - S(\mathbf{t}_i, \mathbf{x}_j))}}{1 + \sum_{j' \in \mathcal{P}_i} e^{\alpha(\lambda - S(\mathbf{t}_i, \mathbf{x}_{j'}))}}, \quad (4)$$

$$\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial S(\mathbf{x}_i, \mathbf{t}_k)} = \frac{\beta}{|\mathcal{N}_i|} \frac{e^{\beta(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda)}}{1 + e^{\beta(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda)}}. \quad (5)$$

As described in [8, 9, 10], the magnitude of Equation 4 is determined by the *relative-hardness*. It means, the gradient is affected by not only the hardness of the similarity to be optimized but also the intra-variance which is expressed as the sum of hardnesses from the positive set. From the softmax-like form, a hard positive ($S < \lambda$) in low intra-variance class has a highlighted gradient, whereas it has a smoothed gradient with others in high intra-variance class.

For the case of Equation 5, the magnitude is determined by the *self-hardness* [9, 10]. As it has the sigmoidal form, the gradient is solely affected by the hardness of the similarity to be optimized. Also, a hard negative ($S > \lambda$) has a not too much gradient as well as its gradient becomes significantly lower after pushed away. This property helps the global structure of embedding space to be converged.

However, while AsyP loss takes advantage of MS loss, BD loss, and the concept of proxy-based DML, it cannot account for the effects of hyper-parameters that vary over training because their values are fixed.

2.2. AdaMS: Adaptive Margin and Scale

Our AdaMS is applied to improve AsyP loss while remaining its inherent advantages. In order to resolve the limitations, we set the following requirements.

- Like the scales α and β , we need two separate margins for positive and negative terms.
- Margins and scales have to be defined for each class since unified values cannot model various data distributions reflecting their intrinsic characteristics.
- Margins and scales should be adjustable according to how well the training optimizes local and global structures on embedding space.

Our main idea satisfying these requirements is to replace the fixed valued hyper-parameters in Equation 2 and Equation 3 with class-dependent learnable parameters for positive and negative terms separately. Similarly to [15, 16, 17], we also employ regularization for margins. Resulting loss terms are given as:

$$\mathcal{L}_i^{\mathcal{P}} = \frac{1}{\text{sg}[\alpha_i]} \log \left(1 + \sum_{j \in \mathcal{P}_i} e^{\alpha_i(\lambda_i^{\mathcal{P}} - S(\mathbf{t}_i, \mathbf{x}_j))} \right) - \omega \lambda_i^{\mathcal{P}}, \quad (6)$$

$$\mathcal{L}_i^{\mathcal{N}} = \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \log \left(1 + e^{\beta_i(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda_i^{\mathcal{N}})} \right) + \omega \lambda_i^{\mathcal{N}}, \quad (7)$$

where $\text{sg}[\cdot]$ is a stop-gradient operation that prevents the gradient in Equation 4 from changing into an unwanted form, ω controls the balance of regularizations, and $\lambda_i^{\mathcal{P}}$, $\lambda_i^{\mathcal{N}}$, α_i , β_i denote the adaptive margins and adaptive scales for class c_i .

2.2.1. Behavior of adaptive margin

To see how the adaptive margins affect the training, we need to analyze the gradients with respect to $\lambda_i^{\mathcal{P}}$ and $\lambda_i^{\mathcal{N}}$ given as:

$$\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial \lambda_i^{\mathcal{P}}} = \frac{\sum_{j \in \mathcal{P}_i} h_{ij}^{\mathcal{P}}}{1 + \sum_{j \in \mathcal{P}_i} h_{ij}^{\mathcal{P}}} - \omega, \quad (8)$$

$$\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial \lambda_i^{\mathcal{N}}} = -\frac{\beta_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \frac{h_{ik}^{\mathcal{N}}}{1 + h_{ik}^{\mathcal{N}}} + \omega, \quad (9)$$

where $h_{ij}^{\mathcal{P}} = e^{\alpha_i(\lambda_i^{\mathcal{P}} - S(\mathbf{t}_i, \mathbf{x}_j))}$ and $h_{ik}^{\mathcal{N}} = e^{\beta_i(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda_i^{\mathcal{N}})}$ denote hardness metrics for class c_i and given similarity. $h_{ij}^{\mathcal{P}}$ and $h_{ik}^{\mathcal{N}}$ have exponentially large values for hard inputs, otherwise they become significantly smaller, close to 0.

If we ignore $-\omega$ in Equation 8, then $\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial \lambda_i^{\mathcal{P}}} > 0$ and its magnitude depends on the number of hard positives. As illustrated in Figure 1.(a)-(d), high intra-variance causes a large gradient, so $\lambda_i^{\mathcal{P}}$ decreases. Then with the more relaxed boundary, some less-hard positives, which were not able to meet the margin before, can come into the area. It leads the model to focus on harder positives. If most positives are well distributed inside the boundary, then the regularization $-\omega$ becomes dominant and $\lambda_i^{\mathcal{P}}$ increases. With the more compact area, hard positives may appear again, then the whole process is repeated.

Likewise, if we ignore ω in Equation 9, then $\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial \lambda_i^{\mathcal{N}}} < 0$ and its magnitude gradually increases when hard negatives exist. As illustrated in Figure 1.(e)-(h), when negatives cross the anchor’s boundary, they cause a large gradient, so $\lambda_i^{\mathcal{N}}$ increases. Then the model narrows the area and pushes the remaining harder negatives outside. If there is no violations, then the regularization ω becomes dominant and $\lambda_i^{\mathcal{N}}$ decreases. By seeking hard negatives again in the broad area, we can achieve higher inter-variance resulting in higher discriminability.

2.2.2. Behavior of adaptive scale

Once the adaptive margin determines the boundary, the adaptive scale adjusts how strictly to punish the violations. Thus their behaviors are highly associated with each other. It can be figured out from the gradients with respect to α_i and β_i :

$$\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial \alpha_i} = \frac{1}{\alpha_i} \frac{\sum_{j \in \mathcal{P}_i} (\lambda_i^{\mathcal{P}} - S(\mathbf{t}_i, \mathbf{x}_j)) h_{ij}^{\mathcal{P}}}{1 + \sum_{j \in \mathcal{P}_i} h_{ij}^{\mathcal{P}}}, \quad (10)$$

$$\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial \beta_i} = \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \frac{(S(\mathbf{x}_i, \mathbf{t}_k) - \lambda_i^{\mathcal{N}}) h_{ik}^{\mathcal{N}}}{1 + h_{ik}^{\mathcal{N}}}. \quad (11)$$

If some positives violate the margin ($S < \lambda_i^{\mathcal{P}}$) resulting in $\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial \alpha_i} > 0$, we can regard the distribution of the class c_i as less-trained yet or inherently difficult to learn. In this case, α_i decreases to smooth the level of $h^{\mathcal{P}}$ and prevent the model from focusing on only a few hard positives. If $\frac{\partial \mathcal{L}_i^{\mathcal{P}}}{\partial \alpha_i} < 0$ that means there are nothing or not many hard positives, then α_i increases to highlight $h^{\mathcal{P}}$ for the outliers and suppress $h^{\mathcal{P}}$ otherwise. These behaviors resemble the property of the relative-hardness of Equation 4, so they jointly boost the discriminative training.

The summation of Equation 11 can be divided into the cases of hard negatives ($S > \lambda_i^{\mathcal{N}}$) and the others. Since $\frac{h^{\mathcal{N}}}{1+h^{\mathcal{N}}} \simeq 1$ for the hard negatives, the sign of the gradient is determined by most negatives outside the anchor’s boundary. If their hardnesses are not low enough for a reason such that the global structure of embedding space approximated by the set of proxies [7] is not constructed yet, then $\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial \beta_i} < 0$ and β_i increases. Also it is accompanied by an increase in $\lambda_i^{\mathcal{N}}$. Thus highly violating proxies are pushed away, which helps establish the global structure. Otherwise, β_i decreases with $\frac{\partial \mathcal{L}_i^{\mathcal{N}}}{\partial \beta_i} > 0$ so that the model can be converged.

2.2.3. Range constraints

It is known that the quality of the resulting embedding space is sensitive to the setting of fixed hyper-parameters [18, 19]. Similarly, if adaptive margins and adaptive scales are implemented without any constraint on their range, they can lead to unstable training. So we impose hyperbolic tangent constraints as:

$$\begin{aligned} \lambda_{i,\text{const}}^{\mathcal{P}} &= \lambda_0 (1 + \tanh(\lambda_i^{\mathcal{P}})) \\ \lambda_{i,\text{const}}^{\mathcal{N}} &= \lambda_0 (1 + \tanh(\lambda_i^{\mathcal{N}})) \\ \alpha_{i,\text{const}} &= \alpha_0 (1 + \delta_\alpha \tanh(\alpha_i)) \\ \beta_{i,\text{const}} &= \beta_0 (1 + \delta_\beta \tanh(\beta_i)) \end{aligned} \quad (12)$$

where δ_α and δ_β control the interval where adaptive scales can vary. Finally, we apply these constrained adaptive margins and constrained adaptive scales to Equation 6 and Equation 7.

3. Experiments

In order to demonstrate the effectiveness of AdaMS, we evaluate the learned AWEs and AGWEs on acoustic word discrimination task and cross-view word discrimination task by using Average Precision (AP) metric [3, 4, 5, 6, 9] in percent (%).

The word-level data is drawn from WSJ dataset [20] with forced alignment; the train/dev/test sets consist of 639501/16839/18274 samples from 13386/3289/3239 unique words. These splits can be obtained through the WSJ recipe officially distributed by the Kaldi toolkit [21]. Then we follow the same feature extraction process as [9].

All methods are implemented with PyTorch [22]. Details that are not mentioned in this paper follow the default setting of PyTorch. For AWEs and AGWEs, we use two 2-layer BLSTM with 512 units per direction as embedding networks (with dropout of 0.4 only for AWEs) and concatenate the last outputs respectively, so $\mathbf{x}, \mathbf{t} \in \mathbb{R}^{1024}$. We use Adam optimizer [23] with learning rate of 10^{-4} for the embedding networks and 10^{-5} for the adaptive margins and adaptive scales.

We set $\lambda = 0.5$, $\alpha = 2$, $\beta = 50$ for the fixed hyper-parameters and for the initial values of the learnable parameters. In the case of using the range constraints as Equation 12, we initialize the adaptive margins and adaptive scales to 0 before applying the constraints. With $\lambda_0 = 0.5$, $\alpha_0 = 2$, $\beta_0 = 50$, $\delta_\alpha = 0.5$, $\delta_\beta = 0.1$, the constrained parameters have the same initial values with non-constrained ones. All these values are chosen by conducting a grid search with the dev set. Note that ω has to differ according to the batch size N . We set $N = 256$, and then $\omega = 0.01$ works well.

3.1. Comparison with other methods

We compare the performance of our AdaMS with other methods including MV Triplet loss [3], MS loss [10], BD loss [11], and the baseline AsyP loss [9]. As originally MS loss and BD loss are not designed for proxy-based DML methods, so we reformulate them as [9]. All experiments are repeated 5 times and we report the mean values with standard deviations.

Table 1: Word discrimination results on WSJ test set. The subscripts indicate the position of proxies: ‘P/N’ for AGWEs as positives and negatives, ‘A’ for AGWEs as anchors.

Methods	Acoustic AP	Cross-view AP
Contrastive Triplet	47.9 (6.86)	-
MV Triplet	81.2 (1.34)	-
MV Triplet	83.3 (0.72)	91.0 (0.56)
Proxy-NCA _{P/N}	86.9 (0.78)	92.7 (0.52)
Proxy-NCA _A	81.5 (0.52)	89.4 (0.40)
Proxy-BD _{P/N}	90.5 (0.40)	95.4 (0.24)
Proxy-BD _A	90.8 (0.52)	95.6 (0.27)
Proxy-MS _{P/N}	90.8 (0.56)	96.3 (0.30)
Proxy-MS _A	88.1 (0.67)	96.4 (0.13)
AsyP	92.1 (0.43)	96.3 (0.23)
AsyP + AdaMS (Ours)	92.7 (0.13)	96.7 (0.12)

The comparison results are summarized in Table 1. It can be seen that our method outperforms all others on both tasks. In [9], AsyP loss improved in Acoustic AP compared to other methods. However, the performance in Cross-view AP was saturated at some level showing little difference with Proxy-MS_{P/N} and Proxy-MS_A losses. When we apply our AdaMS method

to AsyP loss, then we can get extra improvements not only in Acoustic AP but also in Cross-view AP.

3.2. Discriminability of unseen words

In practical situations of open-vocabulary wake-up word detection or query-by-example spoken term detection, which mainly utilizes AWEs, enrolled or queried word is highly probable to have been unseen at the training. Therefore, evaluating the discriminability of unseen words is practically meaningful.

In the test set, there are 402 unseen word classes. To measure the performance in this case, AP for the acoustic word discrimination task is computed by using the samples of unseen words only as the query. The whole test set is used as the retrieval set as the same. To simulate the realistic scenario, the cross-view task is not considered since it is plausible that no text labels are provided for unseen words.

Table 2: Acoustic word discrimination results with unseen word queries on WSJ test set.

Methods	Acoustic AP
AsyP	63.5 (1.72)
AsyP + AdaMS (Ours)	72.8 (2.83)

The results are given in Table 2. The relative improvement about 14.6% clearly demonstrates the outperforming performance of our AdaMS method.

3.3. Ablation study

To investigate the importance of each part of the AdaMS, we conduct an ablation study as shown in Table 3.

Table 3: Ablation study.

Methods	Acoustic AP	Cross-view AP
AsyP (baseline)	92.1 (0.43)	96.3 (0.23)
+ adaptive margin	92.3 (0.37)	96.4 (0.31)
+ range constraints	92.3 (0.22)	96.3 (0.21)
+ adaptive scale	91.2 (0.43)	95.9 (0.22)
+ range constraints	91.7 (0.38)	96.1 (0.21)
+ both	91.8 (0.21)	96.3 (0.11)
+ range constraints (Ours)	92.7 (0.13)	96.7 (0.12)

First, we apply the adaptive margin and adaptive scale separately. Similar to [15, 16, 17], the adaptive margin shows its own effectiveness. However, applying the adaptive scale alone leads to a performance degradation. The reason we can think of is that the decision boundaries are eventually determined by the margins, so if the margins are fixed, then the varying scales rather make the training unstable. Even if we employ both methods, the result gets better but is still lower than the baseline. Here, we can understand that the stability of the training is related to the adaptive scale, and this discovery raises the necessity of introducing the range constraints.

When we use the range constraints to the method applying the adaptive margins alone, there is almost no change. On the other hand, there is an improvement in the adaptive scales with the range constraints. But it seems there is a limit as well since the adaptive margins are not applied together. Finally, when

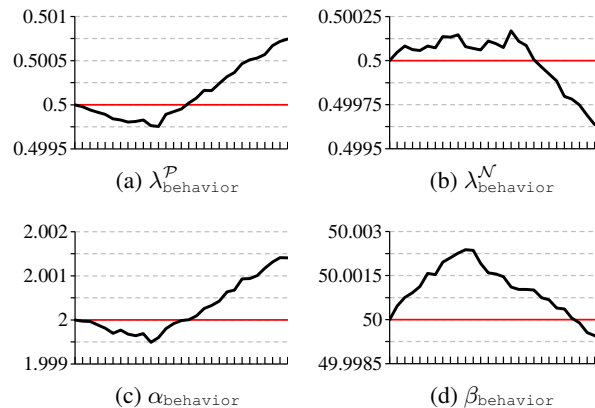


Figure 2: Visualization of how the adaptive margins and adaptive scales for the word ‘behavior’ change during the first epoch. Red horizontal lines indicate the initial values.

we apply the range constraints to the method with the adaptive margins and adaptive scales, we get the best result.

In sum, the class-dependent adaptive margins are effective by themselves, as [15, 16, 17]. To achieve more improvement, we have to apply the adaptive margins and adaptive scales together, but the range constraints are crucial in this case.

3.4. Visualization

In order to see the actual behavior of the adaptive margins and adaptive scales described in Sec.2.2, all values are recorded during the training. Particularly, we focus only on the first epoch. Because all model parameters rapidly move to the nearest optima with the steepest gradients at the beginning, we can observe their distinct behavior.

We select the word class ‘behavior’ and visualize its values in Figure 2. In this example, as we expected, λ^P is highly correlated with α , and they decrease until a particular moment and then increase. Also, though λ^N fluctuates a bit, it shows a similar trend with β , where they increase at first and then decrease. The fluctuation of λ^N can be interpreted that the development of the global structure of the embedding space at the early stage of the training incurs violations and punishing frequently.

4. Conclusion

In this paper, we have proposed a novel AdaMS method to address the problem of fixed hyper-parameters in DML loss functions. It involves simple modification by replacing hyper-parameters of margin and scale with range constrained learnable parameters of adaptive margins and adaptive scales. Since each class can approximate its real distribution closely, we can construct more discriminative embedding space. The outperforming results on word discrimination tasks demonstrate the effectiveness of our method. In addition, through the ablation study, we have demonstrated that applying adaptive margins and adaptive scales together is meaningful and that the range constraints need to be considered at the same time.

5. Acknowledgement

This work is/was supported by SAMSUNG Research, Samsung Electronics Co., Ltd.

6. References

- [1] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4950–4954.
- [2] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 503–510.
- [3] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *Int. Conf. on Learning Representations (ICLR)*, 2017.
- [4] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5641–5645.
- [5] M. Jung, H. Lim, J. Goo, Y. Jung, and H. Kim, "Additional shared decoder on siamese multi-view encoders for learning acoustic word embeddings," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 629–636.
- [6] Y. Hu, S. Settle, and K. Livescu, "Multilingual jointly trained acoustic and written word embeddings," in *Proc. of Ann. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, 2020, pp. 1052–1056.
- [7] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 360–368.
- [8] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3238–3247.
- [9] M. Jung and H. Kim, "Asymmetric proxy loss for multi-view acoustic word embeddings," in *Proc. of Ann. Conf. of the Int. Speech Communication Association (INTERSPEECH)*, 2022, pp. 5170–5174.
- [10] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5022–5030.
- [11] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Int. Conf. on Pattern Recognition*, 2014, pp. 34–39.
- [12] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep metric learning with hierarchical triplet loss," in *Proc. of the Eur. Conf. on Computer Vision (ECCV)*, 2018, pp. 269–285.
- [13] A. Li, W. Huang, X. Lan, J. Feng *et al.*, "Boosting few-shot learning with adaptive margin loss," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 576–12 584.
- [14] Y. Wang, P. Liu, Y. Lang, Q. Zhou, and X. Shan, "Learnable dynamic margin in deep metric learning," *Pattern Recognition*, vol. 132, p. 108961, 2022.
- [15] C. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2840–2848.
- [16] M. Li, S. Zhang, F. Zhu, W. Qian *et al.*, "Symmetric metric learning with adaptive margin for recommendation," *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 04, pp. 4634–4641, 2020.
- [17] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 947–11 956.
- [18] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.
- [19] X. Zhang, R. Zhao, J. Yan, M. Gao *et al.*, "P2SGrad: Refined gradients for optimizing deep face models," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9906–9914.
- [20] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. of the Workshop on Speech and Natural Language*, 1992, pp. 357–362.
- [21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [22] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations (ICLR)*, 2015.