Supplementary Material
Learning Probabilistic Ordinal Embeddings for Uncertainty-Aware Regression

Wanhua Li[1,2], Xiaoke Huang[1,2,3], Jiwen Lu[1,2,*], Jianjiang Feng[1,2], Jie Zhou[1,2]
[1]Department of Automation, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology, China
[3]School of Artificial Intelligence, Beijing Normal University
li-wh17@mails.tsinghua.edu.cn
xiaokehuang@mail.bnu.edu.cn {lujiwen,jfeng,jzhou}@tsinghua.edu.cn

1. Parameters Discussion

In this section, we investigate the influence of different weights of ordinal constraint loss α and VIB loss β on the MORPH II dataset. In previous experiments, we fixed the α to 1e-4, β to 1e-5. To see the effect of these trade-off hyper-parameters, we freeze one parameter as default and tweak the other one. Table 1 shows the comparison results with different α values on the MORPH database. Our method achieves the best performance when α is set to 1e-4. Therefore we set α to 1e-4 in other experiments. Table 2 illustrates the results with varied β values on the MORPH database. The best result is achieved when β is set to 1e-5, which is adopted in the following experiments. We also observe significant performance degradation with large α or β weights (larger than 1e-3), which demonstrates the necessity of hyper-parameter tuning.

Table 1. Results of ordinal loss with different α values on the MORPH II dataset.

α	1e-2	1e-3	1e-4	1e-5	1e-6	1e-7
MAE	3.89	2.42	**2.35**	2.39	2.38	2.40

Table 2. Results of VIB loss with different β values on the MORPH II dataset.

β	1e-2	1e-3	1e-4	1e-5	1e-6	1e-7
MAE	8.36	2.39	2.39	**2.35**	2.37	2.37

2. Distributions of Uncertainty

The harmonic mean of the predicted variance $\boldsymbol{\sigma}$ is employed as the approximated measurement of the estimated uncertainty. To further demonstrate the utility of this metric, We visualize the distributions of the learned uncertainty on the corrupted MORPH II test set using Gaussian blur with three different radii. The results are shown in Figure 1. With different Gaussian blur radii (each equal to 0, 5, 10), the learned uncertainty increases while the image quality degrades. As one can see, the distributions "move" to the right by a large margin as the quality degradation increases in the following order: radius=0 $<$ radius=5 $<$ radius=10.

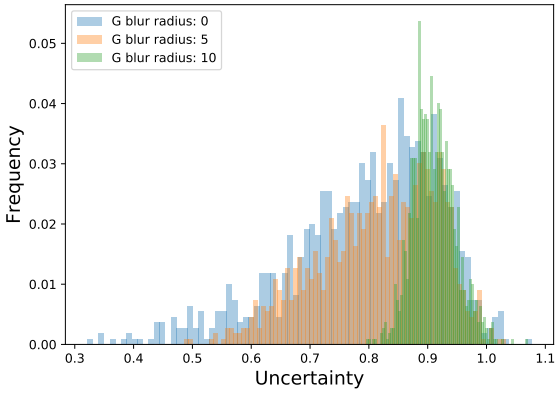

Figure 1. Uncertainty distributions on the corrupted MORPH II test set.

3. Triangle Inequality Issue

The potential problem of symmetric KL divergence is that it is not a strict distance metric due to the triangle inequality issue. However, it is widely used to measure the difference of probability distributions in practice and achieves great success in many areas such as VAE. We also provide the 2-Wasserstein distance in our paper and find that both of them work well. We present more results on other

* Corresponding author

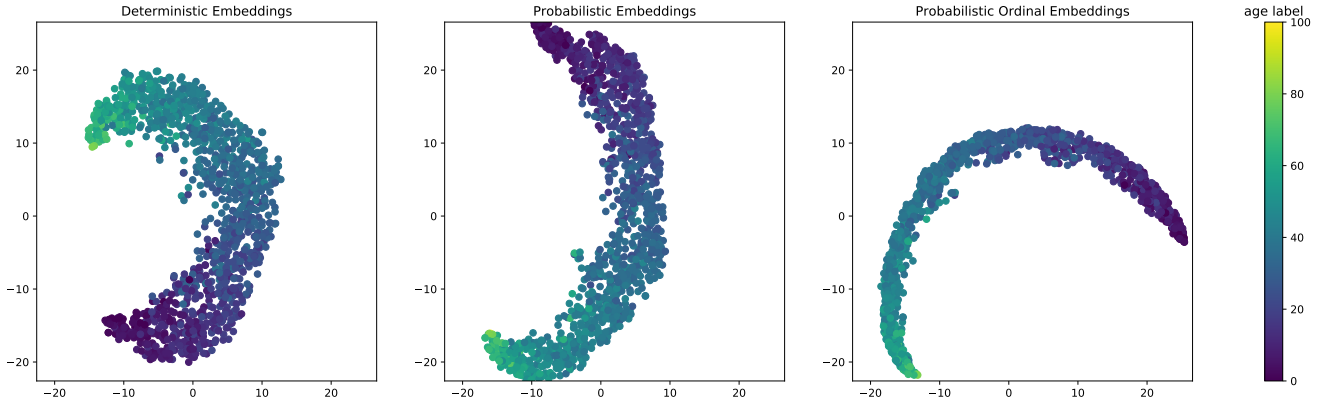

Figure 2. Visualization of feature embeddings with the classification based method on the MORPH II test set.

tasks in Table 3 and observe the same conclusion. Users can choose the suitable metric according to their needs.

Table 3. More results with different metrics. 2-W for 2-Wasserstein distance and S-KL for symmetric KL divergence.

Metric	Adience		Historical	
	Acc (%)	MAE	ACC (%)	MAE
2-W	61.7±4.7	0.45±0.07	51.88±1.98	0.72±0.03
S-KL	60.5±4.4	0.47±0.06	54.68±3.21	0.67±0.04

4. Ordinal Information

To validate that the proposed method can preserve the ordinal information in the embedding space, we conducted quantitative and qualitative evaluations. For quantitative results, we count the proportion of triplets that violate the ordinal constraint in the embedding space on the test set. The results on the MORPH II test set are presented in Table 4. We see that our POE better preserves the ordinal information in the embeddings. For qualitative results, we illustrate the learned features of deterministic embeddings, probabilistic embeddings, and probabilistic ordinal embeddings on the MORPH II test set. Figure 2 shows the results with the classification based method and Figure 3 shows the results with the ranking based method. When using t-SNE, we set the perplexity=100 and fix the initial state. We see that compared with the deterministic embeddings and probabilistic embeddings, POEs learned more compact and ordered feature embeddings, which validates that the ordinal constraint in the target space is well preserved in the learned embedding space.

5. Standard Deviation

Following most previous works, we did not report the standard deviations for the experiments on the MORPH II dataset. Now we list the standard deviations according to the order in Table 3 of our main paper: 0.02, 0.06, 0.02, 0.01, 0.01, 0.03, 0.06, 0.03, 0.02, 0.01. The standard deviations of all six methods in Table 4 of our main paper are as follows: 0.03, 0.01, 0.02, 0.01, 0.03, 0.01. We observe very small standard deviations, indicating that the original conclusions still hold.

Table 4. Quantitative results with different methods. D-E for Deterministic Embeddings, P-E for Probabilistic Embeddings, POE for Probabilistic Ordinal Embeddings (our method).

Method	Classification based			Ranking based		
	D-E	P-E	POE	D-E	P-E	POE
%	28.71	30.44	**15.58**	19.93	18.87	**15.99**

6. The Selection of T

Monte-Carlo sampling is **ONLY** used during **training**. To determine the number T of samples, we show the comparisons of multiply-accumulate operations (MACs) and performance with different sample numbers T in Table 5. We set T = 50 for a good trade-off in our experiments. Note that the extra computation costs are **tiny (∼0.1%)**.

Table 5. The comparison of MACs and performance with different numbers of samples T on the MORPH II dataset. T = 0 indicates deterministic embeddings.

T	0	10	50	100	200
MACs (G)	15.497	15.501	15.517	15.538	15.579
MAE	2.64	2.43	2.35	2.36	2.34

7. Online Hard Example Mining

For a triplet $(\boldsymbol{x}_l, \boldsymbol{x}_m, \boldsymbol{x}_n)$, the relationship is $|\boldsymbol{y}_l - \boldsymbol{y}_m| < |\boldsymbol{y}_l - \boldsymbol{y}_n|$, and we aim to constrain the probabilistic embeddings to satisfy $d(\boldsymbol{z}_l, \boldsymbol{z}_m) < d(\boldsymbol{z}_l, \boldsymbol{z}_n)$. The proposed ordi-

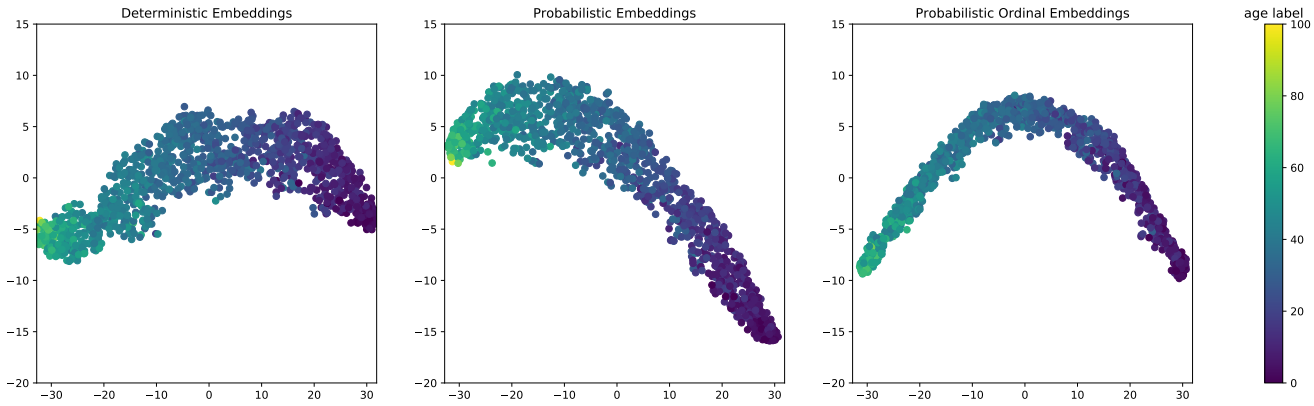

Figure 3. Visualization of feature embeddings with the ranking based method on the MORPH II test set.

nal distribution constraint involves triplet selection. Generating all possible triplets from $\mathcal{S} = \{(l, m, n) \big| |\boldsymbol{y}_l - \boldsymbol{y}_m| < |\boldsymbol{y}_l - \boldsymbol{y}_n|\}$ is inefficient since many of them easily fulfill the ordinal distribution constraint and do not contribute to the network training [1]. One way to address this issue is to select hard triplets. In this paper, we adopt a simple online hard example mining strategy to ensure fast convergence. We assume that the triplet (l, m, n) is a hard one when the difference between $|\boldsymbol{y}_l - \boldsymbol{y}_m|$ and $|\boldsymbol{y}_l - \boldsymbol{y}_n|$ is small. For a batch of training data with N samples, we construct N triplets from them. We first set each sample in the batch as the $anchor$ sample l. Then the next adjacent sample in the batch is selected as the second element of the corresponding triplet. The third element is chosen from the remaining $N-2$ samples which is most likely to violate the constraint according to the above assumption. Specifically, the sample that minimizes $\big| |\boldsymbol{y}_l - \boldsymbol{y}_m| - |\boldsymbol{y}_l - \boldsymbol{y}_n| \big|$ and satisfies $|\boldsymbol{y}_l - \boldsymbol{y}_m| \neq |\boldsymbol{y}_l - \boldsymbol{y}_n|$ is selected. In this way, we obtain N triplets for each batch. Experimental results show that this strategy can achieve satisfactory results. It is an interesting future work to explore more advanced hard example mining strategies.

References

[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 3