Deep Causal Metric Learning

Xiang Deng, Zhongfei Zhang
State University of New York at Binghamton

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Motivation

- In DML, the distance between a pair of images varies with the tasks (i.e., learning goals).
- The background and foreground (i.e., object) in an image can be switched based on the task.
- Backgrounds and objects are typically highly correlated in reality.
- The high correlation between an object and a background makes DML more likely suffer from background (context) biases in the training data, since the classes in the training dataset can be totally different from those in the test dataset in the DML.
- The existing approaches typically focus on designing different hard sample mining or distance margin strategies and then minimize a pair/triplet-based or proxy-based loss over the training data, which can lead the model to recklessly learn all the correlated distances found in training data including the spurious distance that is not the distance of interest.

Figure 1. Biased distance metric induced by context prior.

Contributions

- Different from the existing DML approaches that focus on designing different sampling or distance margin strategies for pair/triplet-based or proxy-based losses, we study DML from a different perspective by proposing deep causal metric learning (DCML) to pursue the true causality of the distances between samples.
- We design a novel metric learning framework, i.e., DCML, that learns the causal distance between samples through explicitly learning context-environment-invariant attention and task-invariant embedding based on causal inference.
- Extensive experiments on several benchmark datasets demonstrate that DCML has a better performance than the existing approaches.

Framework

DCML learns the metric with a de-cofounded model based on backdoor adjustment and invariant risk minimization:

$$\mathcal{L}_{inv} = \sum_{d_j \in D} \left[\mathcal{L}_{env}(d_j, (\mathcal{G}, \mathcal{I}), c) + \alpha * \|\nabla_{c|c=\mathbf{1}} \mathcal{L}_{env}(d_j, (\mathcal{G}, \mathcal{I}), c)\|^2 \right]$$

G and *I* are achieved through environment-invariant attention and task-invariant embedding:

$$\mathcal{L}_{it} = \sum_{d_j \in D} \Big[\mathcal{L}_{env}(d_j, \mathcal{T}_{\theta_j}(h) \circ h, c) + \alpha * \|\nabla_{c|c=\mathbf{1}} \mathcal{L}_{env}(d_j, \mathcal{T}_{\theta_j}(h) \circ h, c)\|^2 + \beta * \|\theta_j - \hat{\theta}\|^2$$
$$+ \gamma * \mathbf{1}_{[y_i == y_k]} * \|\mathcal{T}_{\theta_j}(h_i) \circ h_i - \mathcal{T}_{\theta_j}(h_k) \circ h_k\|^2 \Big]$$

Framework

We also minimize the empirical error on the original dataset which is also an important environment to the task:

$$\mathcal{L} = \mathcal{L}_{it} + \mathcal{L}_{er}(\mathbf{D}, T_\theta(h) \circ h, c)$$

DCML automatically learns the context environments that the current embedding and the attention are not optimal or consistent across:

$$\arg\max_{w} \sum_{d_j \in D} [\|\nabla_{c|c=1} \mathcal{L}_{env}(d_j, \mathcal{T}_{\theta_j}(h) \circ h, c)\|^2 + \|\nabla_{\theta_j|\theta_j=1} \mathcal{L}_{env}(d_j, \mathcal{T}_{\theta_j}(h) \circ h, c)\|^2]$$

Experiments

Table 1. Comparison results (%) on CUB200.

	Concatenated (512-dim)			Separated (128-dim)		
	P@1	RP	MAP@R	P@1	RP	MAP@R
Pretrained	51.05	24.85	14.21	50.54	25.12	14.53
Contrastive	68.13 ± 0.31	37.24 ± 0.28	26.53 ± 0.29	59.73 ± 0.40	31.98 ± 0.29	21.18 ± 0.28
Triplet	64.24 ± 0.26	34.55 ± 0.24	23.69 ± 0.23	55.76 ± 0.27	29.55 ± 0.16	18.75 ± 0.15
NT-Xent	66.61 ± 0.29	35.96 ± 0.21	25.09 ± 0.22	58.12 ± 0.23	30.81 ± 0.17	19.87 ± 0.16
ProxyNCA	65.69 ± 0.43	35.14 ± 0.26	24.21 ± 0.27	57.88 ± 0.30	30.16 ± 0.22	19.32 ± 0.21
Margin	63.60 ± 0.48	33.94 ± 0.27	23.09 ± 0.27	54.78 ± 0.30	28.86 ± 0.18	18.11 ± 0.17
Margin/class	64.37 ± 0.18	34.59 ± 0.16	23.71 ± 0.16	55.56 ± 0.16	29.32 ± 0.15	18.51 ± 0.13
N. Softmax	65.65 ± 0.30	35.99 ± 0.15	25.25 ± 0.13	58.75 ± 0.19	31.75 ± 0.12	20.96 ± 0.11
COS	67.32 ± 0.32	37.49 ± 0.21	26.70 ± 0.23	59.63 ± 0.36	31.99 ± 0.22	21.21 ± 0.22
ArcFace	67.50 ± 0.25	37.31 ± 0.21	26.45 ± 0.20	60.17 ± 0.32	32.37 ± 0.17	21.49 ± 0.16
FastAP	63.17 ± 0.34	34.20 ± 0.20	23.53 ± 0.20	55.58 ± 0.31	29.72 ± 0.16	19.09 ± 0.16
SNR	66.44 ± 0.56	36.56 ± 0.34	25.75 ± 0.36	58.06 ± 0.39	31.21 ± 0.28	20.43 ± 0.28
MS	65.04 ± 0.28	35.40 ± 0.12	24.70 ± 0.13	57.60 ± 0.24	30.84 ± 0.13	20.15 ± 0.14
MS+Miner	67.73 ± 0.18	37.37 ± 0.19	26.52 ± 0.18	59.41 ± 0.30	31.93 ± 0.15	21.01 ± 0.14
SoftTriple	67.27 ± 0.39	37.34 ± 0.19	26.51 ± 0.20	59.94 ± 0.33	32.12 ± 0.14	21.31 ± 0.14
ProxyNCA++	64.69 ± 0.40	34.37 ± 0.13	23.53 ± 0.12	57.13 ± 0.36	29.52 ± 0.16	18.76 ± 0.15
ContXBM	68.43 ± 1.18	37.66 ± 0.56	26.85 ± 0.63	60.95 ± 0.76	32.69 ± 0.33	21.78 ± 0.35
Proxy-Anchor	67.64 ± 0.42	37.29 ± 0.19	26.47 ± 0.21	60.59 ± 0.24	32.45 ± 0.15	21.57 ± 0.15
DCML (Ours)	**70.09 ± 0.22**	**39.05 ± 0.13**	**28.36 ± 0.13**	**62.28 ± 0.30**	**33.39 ± 0.18**	**22.61 ± 0.15**

Experiments

Table 2. Comparison results (%) on Car-196.

	Concatenated (512-dim)			Separated (128-dim)		
	P@1	RP	MAP@R	P@1	RP	MAP@R
Pretrained	46.89	13.77	5.91	43.27	13.37	5.64
Contrastive	81.78 ± 0.43	35.11 ± 0.45	24.89 ± 0.50	69.80 ± 0.38	27.78 ± 0.34	17.24 ± 0.35
Triplet	79.13 ± 0.42	33.71 ± 0.45	23.02 ± 0.51	65.68 ± 0.58	26.67 ± 0.36	15.82 ± 0.36
NT-Xent	80.99 ± 0.54	34.96 ± 0.38	24.40 ± 0.41	68.16 ± 0.36	27.66 ± 0.23	16.78 ± 0.24
ProxyNCA	83.56 ± 0.27	35.62 ± 0.28	25.38 ± 0.31	73.46 ± 0.23	28.90 ± 0.22	18.29 ± 0.22
Margin	81.16 ± 0.50	34.82 ± 0.31	24.21 ± 0.34	68.24 ± 0.35	27.25 ± 0.19	16.40 ± 0.20
Margin/class	80.04 ± 0.61	33.78 ± 0.51	23.11 ± 0.55	67.54 ± 0.60	26.68 ± 0.40	15.88 ± 0.39
N. Softmax	83.16 ± 0.25	36.20 ± 0.26	26.00 ± 0.30	72.55 ± 0.18	29.35 ± 0.20	18.73 ± 0.20
COS	85.52 ± 0.24	37.32 ± 0.28	27.57 ± 0.30	74.67 ± 0.20	29.01 ± 0.11	18.80 ± 0.12
ArcFace	85.44 ± 0.28	37.02 ± 0.29	27.22 ± 0.30	72.10 ± 0.37	27.29 ± 0.17	17.11 ± 0.18
FastAP	78.45 ± 0.52	33.61 ± 0.54	23.14 ± 0.56	65.08 ± 0.36	26.59 ± 0.36	15.94 ± 0.34
SNR	82.02 ± 0.48	35.22 ± 0.43	25.03 ± 0.48	69.69 ± 0.46	27.55 ± 0.25	17.13 ± 0.26
MS	85.14 ± 0.29	38.09 ± 0.19	28.07 ± 0.22	73.77 ± 0.19	29.92 ± 0.16	19.32 ± 0.18
MS+Miner	83.67 ± 0.34	37.08 ± 0.31	27.01 ± 0.35	71.80 ± 0.22	29.44 ± 0.21	18.86 ± 0.20
SoftTriple	84.49 ± 0.26	37.03 ± 0.21	27.08 ± 0.21	73.69 ± 0.21	29.29 ± 0.16	18.89 ± 0.16
ProxyNCA++	82.09 ± 0.41	36.31 ± 0.24	26.02 ± 0.26	70.60 ± 0.18	29.35 ± 0.08	18.63 ± 0.09
ContXBM	83.67 ± 0.35	36.10 ± 0.19	26.04 ± 0.24	72.58 ± 0.21	28.55 ± 0.10	18.07 ± 0.11
Proxy-Anchor	86.38 ± 0.15	37.53 ± 0.17	27.77 ± 0.20	76.85 ± 0.13	30.12 ± 0.10	19.82 ± 0.10
DCML (Ours)	**87.43 ± 0.21**	**39.60 ± 0.16**	**30.29 ± 0.12**	**78.58 ± 0.27**	**31.58 ± 0.15**	**21.55 ± 0.14**

Experiments

Table 3. Comparison results (%) on SOP.

	Concatenated (512-dim)			Separated (128-dim)		
	P@1	RP	MAP@R	P@1	RP	MAP@R
Pretrained	50.71	25.97	23.44	47.25	23.84	21.36
Contrastive	73.12 ± 0.20	47.29 ± 0.24	44.39 ± 0.24	69.34 ± 0.26	43.41 ± 0.28	40.37 ± 0.28
Triplet	72.65 ± 0.28	46.46 ± 0.38	43.37 ± 0.37	67.33 ± 0.34	40.94 ± 0.39	37.70 ± 0.38
NT-Xent	74.22 ± 0.22	48.35 ± 0.26	45.31 ± 0.25	69.88 ± 0.19	43.51 ± 0.21	40.31 ± 0.20
ProxyNCA	75.89 ± 0.17	50.10 ± 0.22	47.22 ± 0.21	71.30 ± 0.20	44.71 ± 0.21	41.74 ± 0.21
Margin	70.99 ± 0.36	44.94 ± 0.43	41.82 ± 0.43	65.78 ± 0.34	39.71 ± 0.40	36.47 ± 0.39
N. Softmax	75.36 ± 0.17	50.01 ± 0.22	47.13 ± 0.22	71.65 ± 0.14	45.32 ± 0.17	42.35 ± 0.16
COS	75.79 ± 0.14	49.77 ± 0.19	46.92 ± 0.19	70.71 ± 0.19	43.56 ± 0.21	40.69 ± 0.21
ArcFace	76.20 ± 0.27	50.27 ± 0.38	47.41 ± 0.40	70.88 ± 1.51	44.00 ± 1.26	41.11 ± 0.22
FastAP	72.59 ± 0.26	46.60 ± 0.29	43.57 ± 0.28	68.13 ± 0.25	42.06 ± 0.25	38.88 ± 0.25
SNR	73.40 ± 0.09	47.43 ± 0.13	44.54 ± 0.13	69.45 ± 0.10	43.34 ± 0.12	40.31 ± 0.12
MS	74.50 ± 0.24	48.77 ± 0.32	45.79 ± 0.32	70.43 ± 0.33	44.25 ± 0.38	41.15 ± 0.38
MS+Miner	75.09 ± 0.17	49.51 ± 0.20	46.55 ± 0.20	71.25 ± 0.15	45.19 ± 0.16	42.10 ± 0.16
SoftTriple	76.12 ± 0.17	50.21 ± 0.18	47.35 ± 0.19	70.88 ± 0.20	43.83 ± 0.20	40.92 ± 0.20
ProxyNCA++	75.10 ± 0.15	49.50 ± 0.19	46.56 ± 0.19	70.43 ± 0.17	43.82 ± 0.20	41.51 ± 0.18
Proxy-Anchor	76.12 ± 0.19	50.82 ± 0.27	47.88 ± 0.26	72.79 ± 0.22	47.00 ± 0.24	43.97 ± 0.25
DCML (Ours)	**77.88 ± 0.19**	**52.81 ± 0.22**	**50.00 ± 0.22**	**73.83 ± 0.21**	**47.38 ± 0.23**	**44.52 ± 0.22**

Conclusion

- In this paper, we study deep metric learning from a novel perspective and accordingly propose deep causal metric learning.
- DCML learns the causal distance metric regarding a task by removing the effects of the spurious distances. This is achieved by learning environment-invariant attention and task-invariant embedding.
- Extensive experiments on several metric learning benchmark datasets demonstrate the effectiveness and superiority of DCML.

Reference

Please refer to the Reference section in "Deep Causal Metric Learning, ICML'2022".

Thank you!