Asymmetric metric learning for knowledge transfer

Mateusz Budnik, Yannis Avrithis

Inria, CNRS, Univ Rennes, IRISA

Motivation

- Given a database of images indexed by a large DL model
- Queries from small network (e.g. working on a mobile device)
- Re-index the whole database by the small model for it to match?
- Or: can the small model learn to directly map queries like the large model

Contributions

- Knowledge transfer for pair-based metric learning for instance-level image retrieval
- Introduction of asymmetric metric learning paradigm
- Asymmetric testing: database represented by large network, queries by lightweight network on device, no re-indexing

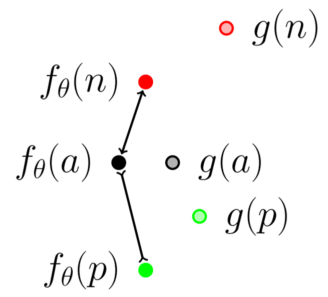
Metric learning

- *Cosine similarity* is used in this work
- The *symmetric similarity* $s_\theta^{\text{sym}}(a,x)$ between an anchor $a \in X$ and a positive or negative example $x \in P(a) \cup N(a)$ is obtained by representing both in the feature space of the student f_θ:
$$s_\theta^{\text{sym}}(a,x) := \text{sim}(f_\theta(a), f_\theta(x)).$$
- This is a standard setting for metric learning
- Symmetric testing is based on this similarity

Asymmetric metric learning (AML)

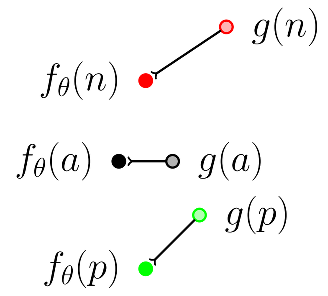
- Instance-level image retrieval
- Applies to both symmetric and asymmetric testing
- Combines knowledge transfer with supervised metric learning
- In *asymmetric similarity* $s_\theta^{\text{asym}}(a,x)$ the anchor a is represented by the student f_θ, while positive and negative examples x are represented by the teacher g:
$$s_\theta^{\text{asym}}(a,x) := \text{sim}(f_\theta(a), g(x))$$
- Can be used with any supervised metric learning loss

Metric learning and knowledge transfer



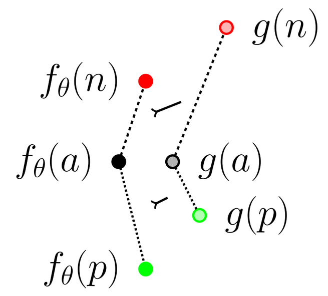
Symmetric

- labels used, teacher not used (f_θ: student, g: teacher)
- positive pairs mutually attract and negative pairs repulse in student space
- losses: Contr, Triplet, MS [6]



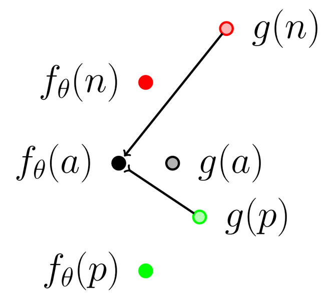
Regression

- labels not used, teacher used
- examples in student space attracted to same examples in teacher space
- loss: Reg



Relational

- labels not used, teacher used
- pairwise relations encouraged to be compatible in both spaces
- losses: RKD [2], DR [1]



Asymmetric

- both labels and teacher used
- positives attract, negatives repulse using asymmetric similarity
- losses: any supervised losses

Best loss functions

- Regression (Reg)
$$\ell_{\text{R}}(a;\theta) := -s_\theta^{\text{asym}}(a,a) = -\text{sim}(f_\theta(a), g(a))$$
- Asymmetric contrastive (Contr)
$$\ell_{\text{C}}(a;\theta) := \sum_{n \in N(a)} [s_\theta^{\text{asym}}(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta^{\text{asym}}(a,p)$$
- Asymmetric contrastive + regression (Contr$^+$)
$$\ell_{\text{C}^+}(a;\theta) := \sum_{n \in N(a)} [s_\theta^{\text{asym}}(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta^{\text{asym}}(a,p) \quad - s_\theta^{\text{asym}}(a,a)$$

Experimental setup

- Training set: SfM120k [4]
  - Positives: based on closest camera position to query and number of inliers
  - Negatives: k-nearest neighbours from non-matching clusters
- Test sets: revisited Oxford and Paris [3] (including results with 1M distractors)

Network	Teacher	d	GFLOPS	Param(M)
ResNet101		2048	42.85	42.50
EfficientNet-B3		1536	5.36	10.70
	ResNet101	2048	6.26	13.84

- teacher: ResNet101 (RN101)
- student: EfficientNet-B3[5] (EN-B3), $7\times$ less FLOPS and $3\times$ less parameters

Symmetric testing

Stu	d	Tea	Lab	Mining	Asym	Loss	Medium $\mathcal{R}$Oxf	Medium $\mathcal{R}$Par	Hard $\mathcal{R}$Oxf	Hard $\mathcal{R}$Par
RN101	2048		✓	hard		Contr	65.4	76.7	40.1	55.2
EN-B3	512		✓	hard		Contr	53.8	70.9	26.2	46.0
	2048		✓	hard		Contr	59.6	75.1	33.3	51.9
			✓	hard	✓	Contr$^+$	**66.8**	77.1	**42.5**	**55.5**
			✓	hard	✓	Contr	66.3	**77.4**	41.3	**55.5**
			✓	hard	✓	Triplet	39.5	69.4	11.6	45.8
EN-B3	2048	RN101	✓	hard	✓	MS	39.9	69.7	11.7	46.2
				–	✓	Reg	64.9	74.4	40.5	52.4
				random		RKD	56.3	73.0	30.5	50.4
				random		DR	40.3	69.9	11.8	46.4

- Performance measured by mAP
- Contr and Contr$^+$: student beats teacher
- Reg: second best, slightly below teacher
- Everything else worse than student alone

Asymmetric testing

Stu	d	Tea	Lab	Mining	Asym	Loss	Medium $\mathcal{R}$Oxf	Medium $\mathcal{R}$Par	Hard $\mathcal{R}$Oxf	Hard $\mathcal{R}$Par
RN101	2048		✓	hard		Contr	65.4	76.7	40.1	55.2
EN-B3	512		✓	hard		Contr	53.8	70.9	26.2	46.0
	2048		✓	hard		Contr	59.6	75.1	33.3	51.9
			✓	hard	✓	Contr$^+$	45.2	63.7	19.6	40.9
			✓	hard	✓	Contr	37.4	57.4	10.9	33.7
			✓	hard	✓	Triplet	1.5	4.0	0.7	2.5
EN-B3	2048	RN101	✓	hard	✓	MS	1.5	4.0	0.7	2.4
				–	✓	Reg	**52.9**	**65.2**	**27.8**	**42.4**
				random		RKD	1.6	3.8	0.7	2.4
				random		DR	1.5	4.0	0.7	2.5

- Reg: best, but significantly lower than student alone
- Contr$^+$/ Contr: second / third best, significantly lower than Reg
- RKD, DR: completely fail (expected, absolute coordinates needed)
- Triplet, MS: completely fail (unexpected)

References

[1] Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[2] W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[3] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*. Springer, 2016.

[5] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 2019.

[6] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

Contact: yannis@avrithis.com