Metric learning: Knowledge transfer, data augmentation, and attention

Yannis Avrithis

Athena Research Center

ICCV 2021 Tutorial: **Large-Scale Visual Localization**
Virtual, October 2021

ATHENA Research & Innovation Information Technologies

2021 ICCV OCTOBER 11-17 VIRTUAL

context

- representation learning for instance-level tasks like visual localization often reduces to metric learning
- ideas addressed most commonly in classification, less so in metric learning
  - knowedge transfer (from teacher to student models)
  - data augmentation (mixup)
  - attention (channel/spatial, local/global)

knowledge transfer

asymmetric metric learning for knowledge transfer

[CVPR 2021]

Mateusz Budnik

Yannis Avrithis

paper

https://arxiv.org/abs/2006.16331

code

https://github.com/budnikm/asymmetric_metric_learning

asymmetric metric learning (AML)

- instance-level image retrieval
- asymmetric testing: database represented by large network, queries by lightweight network on device, no re-indexing
- asymmetric metric learning: use asymmetric representations at training in teacher-student setup
- applies to both symmetric and asymmetric testing
- combines of knowledge transfer with supervised metric learning

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

asymmetric metric learning (AML)

- instance-level image retrieval
- asymmetric testing: database represented by large network, queries by lightweight network on device, no re-indexing
- asymmetric metric learning: use asymmetric representations at training in teacher-student setup
- applies to both symmetric and asymmetric testing
- combines of knowledge transfer with supervised metric learning

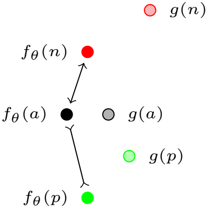
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

asymmetric metric learning (AML)

- instance-level image retrieval
- asymmetric testing: database represented by large network, queries by lightweight network on device, no re-indexing
- asymmetric metric learning: use asymmetric representations at training in teacher-student setup
- applies to both symmetric and asymmetric testing
- combines of knowledge transfer with supervised metric learning

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer

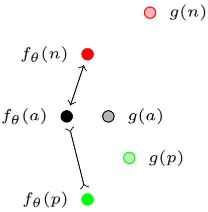


symmetric

- labels used, teacher not used
- positive pairs of examples mutually attracted and negative pairs are repulsed in student space

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



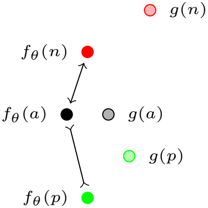
symmetric

- labels used, teacher not used (f_θ: student, g: teacher)
- contrastive $\ell_{\mathrm{C}}(a;\theta)$: independently, positive examples p close to anchor a, negative n farther from a by margin m in student space

$$\sum_{p \in P(a)} -s_\theta(a,p) + \sum_{n \in N(a)} [s_\theta(a,n) - m]_+$$

Hadsell, Chopra, Lecun. CVPR 2006. Dimensionality reduction by learning an invariant mapping.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



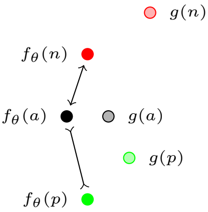
symmetric

- labels used, teacher not used (f_θ: student, g: teacher)
- triplet $\ell_{\mathrm{T}}(a;\theta)$: positive examples p closer to the anchor a than negative n by margin m in student space

$$\sum_{(p,n)\in L(a)} [s_\theta(a,n) - s_\theta(a,p) + m]_+$$

Wang, Song, Leung, Rosenberg, Wang, Philbin, Chen, Wu. CVPR 2014. Learning fine-grained image similarity with deep ranking.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



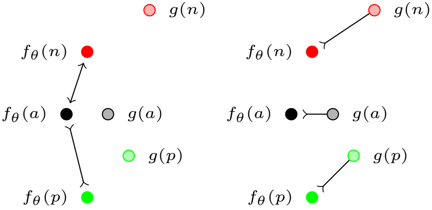
symmetric

- labels used, teacher not used (f_θ: student, g: teacher)
- multi-similarity $\ell_{\mathrm{MS}}(a;\theta)$: positives p (negatives n) farthest from (nearest to) anchor a receive the greatest relative weight

$$\frac{1}{\alpha}\log\left(1+\sum_{p\in P(a)} e^{-\alpha(s_\theta(a,p)-m)}\right)+\frac{1}{\beta}\log\left(1+\sum_{n\in N(a)} e^{\beta(s_\theta(a,n)-m)}\right)$$

Wang, Han, Huang, Dong, Scott. CVPR 2019. Multi-similarity loss with general pair weighting for deep metric learning.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer

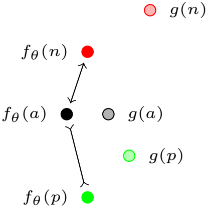


symmetric **regression**

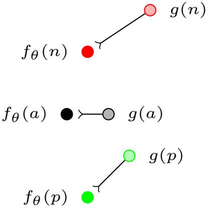
- labels not used, teacher used
- examples in student space attracted to corresponding examples in teacher space

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



symmetric



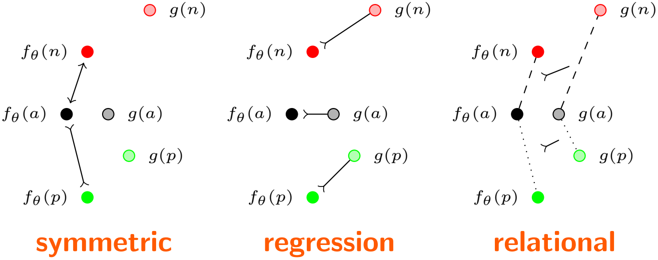
regression

- labels not used, teacher used (f_θ: student, g: teacher)
- regression $\ell_{\mathrm{R}}(a;\theta)$: representations of same example a by two models f_θ, g close to each other, where g is fixed

$$-s_\theta^{\mathrm{asym}}(a,a) = -\operatorname{sim}(f_\theta(a), g(a))$$

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

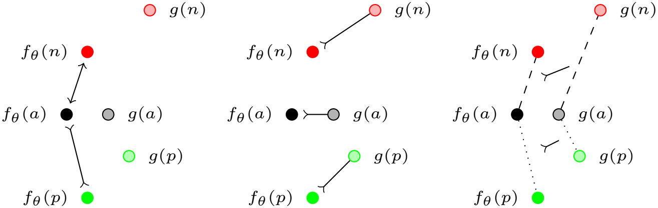
metric learning and knowledge transfer



- labels not used, teacher used
- pairwise / groupwise relations like distances, angles or ranks encouraged to be compatible in both spaces

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



symmetric **regression** **relational**

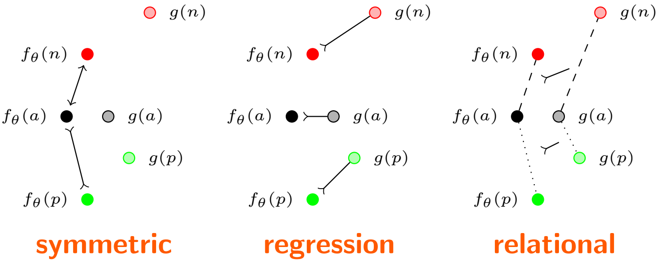
- labels not used, teacher used (f_θ: student, g: teacher)
- relational distillation $\ell_{\mathrm{RKD}}(a;\theta)$: measurements $\psi(\mathbf{a}, \mathbf{x}, \dots)$ of same examples $(a, x, \dots)$ by two models f_θ, g close to each other

$$\sum_{(x,\dots)\in U(a)^n} -\operatorname{sim}(\psi(f_\theta(a), f_\theta(x), \dots), \psi(g(a), g(x), \dots))$$

e.g. distance $\|\mathbf{a} - \mathbf{x}\|$, angle $\operatorname{sim}(\mathbf{a} - \mathbf{x}, \mathbf{a} - \mathbf{y})$; regression $\psi(\mathbf{a}) := \mathbf{a}$

Park, Kim, Lu, Cho. CVPR 2019. Relational knowledge distillation.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer

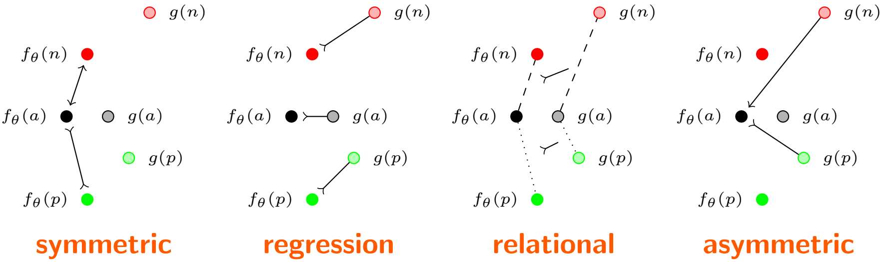


- labels not used, teacher used (f_θ: student, g: teacher)
- DarkRank $\ell_{\mathrm{DR}}(a;\theta)$: examples $y \in V(a,x)$ mapped farther from anchor a than x in teacher space do the same in student space:

$$-\sum_{x \in U(a)} \left(s_\theta^{\mathrm{sym}}(a,x) - \log \sum_{y \in V(a,x)} e^{s_\theta^{\mathrm{sym}}(a,y)} \right)$$

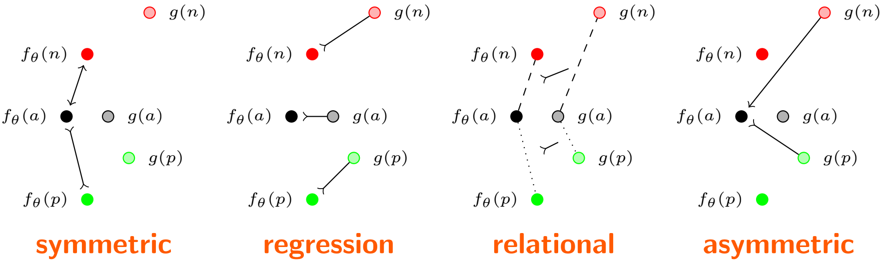
Chen, Wang, Zhang. AAAI 2018. DarkRank: Accelerating deep metric learning via cross sample similarities transfer.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

metric learning and knowledge transfer



- both labels and teacher used
- anchors in student space attracted to positives and repulsed from negatives in teacher space

metric learning and knowledge transfer



- both labels and teacher used (f_θ: student, g: teacher)
- Asymmetric Metric Learning (AML): simply use

$$s_\theta^{\mathrm{asym}}(a, x) := \mathrm{sim}(f_\theta(a), g(x))$$

with any supervised metric learning loss like ℓ_{C}, ℓ_{T}, ℓ_{MS}

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

best loss functions

- regression (Reg)

$$\ell_{\mathrm{R}}(a;\theta) := -s_\theta^{\mathrm{asym}}(a,a) = -\operatorname{sim}(f_\theta(a), g(a))$$

- asymmetric contrastive (Contr)

$$\ell_{\mathrm{C}}(a;\theta) := \sum_{n \in N(a)} [s_\theta(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta(a,p)$$

- asymmetric contrastive + regression (Contr$^+$)

$$\ell_{\mathrm{C}^+}(a;\theta) := \sum_{n \in N(a)} [s_\theta(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta(a,p) \quad - s_\theta(a,a)$$

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

best loss functions

- regression (Reg)

$$\ell_{\mathrm{R}}(a;\theta) := \boxed{-s_\theta^{\mathrm{asym}}(a,a)} = -\operatorname{sim}(f_\theta(a), g(a))$$

- asymmetric contrastive (Contr)

$$\ell_{\mathrm{C}}(a;\theta) := \boxed{\sum_{n \in N(a)} [s_\theta(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta(a,p)}$$

- asymmetric contrastive + regression (Contr$^+$)

$$\ell_{\mathrm{C}^+}(a;\theta) := \boxed{\sum_{n \in N(a)} [s_\theta(a,n) - m]_+ - \sum_{p \in P(a)} s_\theta(a,p)} \boxed{- s_\theta(a,a)}$$

test set: revisited Oxford and Paris

- 11 + 11 landmarks, 70 + 70 queries, 5k + 6k images, easy/hard
- 1M distractor images
- performance measured by mAP: positive ranked first

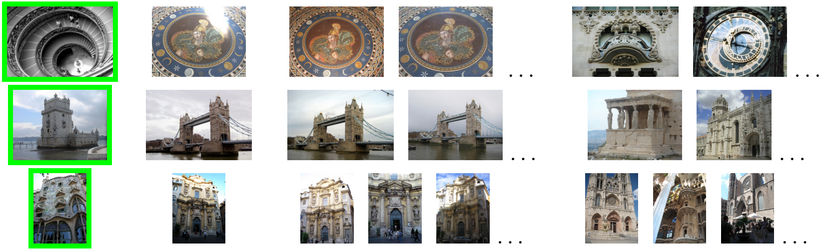
Radenovic, Iscen, Tolias, Avrithis, Chum. CVPR 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

training set: SfM120k (positives)

- camera position (closest to query)
- number of inliers (co-observed 3D points with query)
- according to SIFT descriptors

Radenovic, Tolias, Chum. ECCV 2016. CNN Image Retrieval Learns From BoW: Unsupervised Fine-Tuning with Hard Examples.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

training set: SfM120k (negatives)



- k-nearest neighbors from non-matching clusters
- at most one image per cluster
- according to learned descriptors

Radenovic, Tolias, Chum. ECCV 2016. CNN Image Retrieval Learns From BoW: Unsupervised Fine-Tuning with Hard Examples.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

network models

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 (EN-B3), dimensions d adapted to teacher
- $7\times$ less FLOPS
- $3\times$ less parameters

Tan and Le. ICML 2019. EfficientNet: Rethinking model scaling for convolutional neural networks.
Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr^+: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr^{+}: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr$^+$: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr$^+$: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr^{+}: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr$^+$: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr^{+}: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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

- Contr, Contr$^+$: student beats teacher
- Reg: second best, slightly below teacher
- everything else fails (worse than student alone)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
EN-B3	2048	RN101	✓	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
			✓	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
-

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
EN-B3	2048	RN101	✓	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
			✓	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

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
EN-B3	2048	RN101	✓	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
			✓	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
-

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

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
-

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
EN-B3	2048	RN101	✓	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
			✓	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

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
EN-B3	2048	RN101	✓	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
			✓	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)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

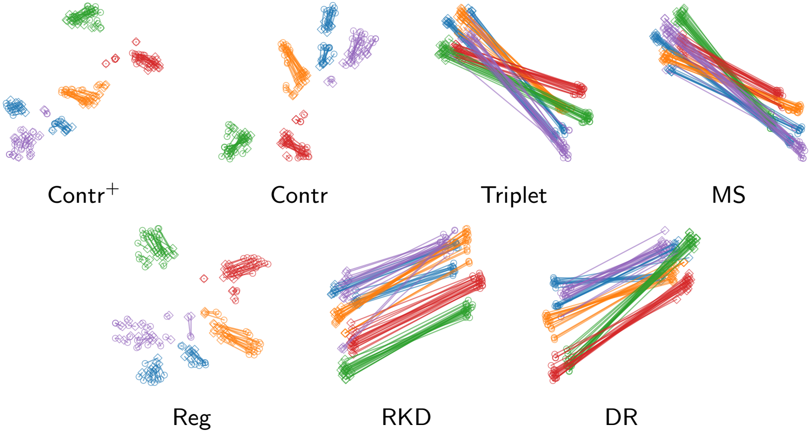
asymmetric testing

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
EN-B3	2048	RN101	✓	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
			✓	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
- Triplet, MS: completely fail (unexpected)

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

asymmetric testing: T-SNE embeddings



- 5 Oxford classes, 20 “easy” examples per class
- Triplet, MS, RKD, DR fail completely

Budnik and Avrithis. CVPR 2021. Asymmetric Metric Learning for Knowledge Transfer.

data augmentation

mixup for deep metric learning

Shashanka Venkataramanan

Ewa Kijak

Laurent Amsaleg

Yannis Avrithis

paper

https://arxiv.org/abs/2106.04990

code

upon publication

data augmentation and mixup

- data augmentation increases the amount and diversity of data, improving the generalization performance at almost no cost
- operates on one image at a time, limited to label-preserving transformations: hard to explore beyond the image manifold
- mixup operates on two or more examples at a time, interpolating examples and labels
- in classification, smooths decision boundaries far away from training data and reduces overly confident predictions
- how about metric learning?

data augmentation and mixup

- data augmentation increases the amount and diversity of data, improving the generalization performance at almost no cost
- operates on one image at a time, limited to label-preserving transformations: hard to explore beyond the image manifold
- mixup operates on two or more examples at a time, interpolating examples and labels
- in classification, smooths decision boundaries far away from training data and reduces overly confident predictions
- how about metric learning?

input mixup and manifold mixup

- standard mixup operation: linear interpolation

$$\text{mix}_\lambda(x, x') := \lambda x + (1 - \lambda)x'$$

  where $\lambda \in [0, 1]$: interpolation factor, drawn from Beta distribution

- interpolation of examples: decomposing model as $f = f_m \circ g_m$,

$$f_\lambda(x, x') := \begin{cases} f(\text{mix}_\lambda(x, x')), & \text{input mixup} \\ f_m(\text{mix}_\lambda(g_m(x), g_m(x'))), & \text{feature mixup} \\ \text{mix}_\lambda(f(x), f(x')), & \text{embedding mixup} \end{cases}$$

- interpolation of labels: $\text{mix}_\lambda(y, y')$
- classification: one-hot encoded class label $y \in \{0, 1\}^C$ per example
- metric learning: labels refer to pairs of examples

Zhang, Cisse, Dauphin and Lopez-Paz. ICLR 2018. mixup: Beyond empirical risk minimization.
Verma, Lamb, Beckham, Najafi, Mitliagkas, Lopez-Paz and Bengio. ICML 2019. Manifold mixup: Better representations by interpolating hidden states.

input mixup and manifold mixup

- standard mixup operation: linear interpolation

$$\mathrm{mix}_\lambda(x, x') := \lambda x + (1-\lambda) x'$$

where $\lambda \in [0,1]$: interpolation factor, drawn from Beta distribution

- interpolation of examples: decomposing model as $f = f_m \circ g_m$,

$$f_\lambda(x, x') := \begin{cases} f(\mathrm{mix}_\lambda(x, x')), & \text{input mixup} \\ f_m(\mathrm{mix}_\lambda(g_m(x), g_m(x'))), & \text{feature mixup} \\ \mathrm{mix}_\lambda(f(x), f(x')), & \text{embedding mixup} \end{cases}$$

- interpolation of labels: $\mathrm{mix}_\lambda(y, y')$
- classification: one-hot encoded class label $y \in \{0,1\}^C$ per example
- metric learning: labels refer to pairs of examples

Zhang, Cisse, Dauphin and Lopez-Paz. ICLR 2018. mixup: Beyond empirical risk minimization.
Verma, Lamb, Beckham, Najafi, Mitliagkas, Lopez-Paz and Bengio. ICML 2019. Manifold mixup: Better representations by interpolating hidden states.

input mixup and manifold mixup

- standard mixup operation: linear interpolation

$$\mathrm{mix}_\lambda(x, x') := \lambda x + (1-\lambda) x'$$

where $\lambda \in [0,1]$: interpolation factor, drawn from Beta distribution

- interpolation of examples: decomposing model as $f = f_m \circ g_m$,

$$f_\lambda(x, x') := \begin{cases} f(\mathrm{mix}_\lambda(x, x')), & \text{input mixup} \\ f_m(\mathrm{mix}_\lambda(g_m(x), g_m(x'))), & \text{feature mixup} \\ \mathrm{mix}_\lambda(f(x), f(x')), & \text{embedding mixup} \end{cases}$$

- interpolation of labels: $\mathrm{mix}_\lambda(y, y')$
- classification: one-hot encoded class label $y \in \{0,1\}^C$ per example
- metric learning: labels refer to pairs of examples

Zhang, Cisse, Dauphin and Lopez-Paz. ICLR 2018. mixup: Beyond empirical risk minimization.
Verma, Lamb, Beckham, Najafi, Mitliagkas, Lopez-Paz and Bengio. ICML 2019. Manifold mixup: Better representations by interpolating hidden states.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels > 1	Labels Mix	Anc-Neg
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels		Anc-Neg
					>1	Mix	
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels > 1	Labels Mix	Anc-Neg
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels > 1	Labels Mix	Anc-Neg
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels		Anc-Neg
					> 1	Mix	
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels > 1	Labels Mix	Anc-Neg
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

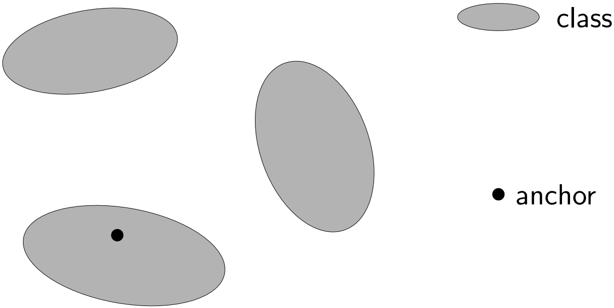
Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

existing approaches

Method	DML	Stoch	Pairs	Proxy	Labels >1	Labels Mix	Anc-Neg
Hardness-Aware DML	✓		✓				
Embedding Expansion	✓		✓				
Symmetrical Synthesis	✓		✓				
Proxy Synthesis	✓	✓		✓	✓		✓
MoCHi		✓	✓		✓		✓
i-Mix		✓	✓		✓	✓	
MixCo		✓	✓		✓	✓	
Metrix (ours)	✓	✓	✓	✓	✓	✓	✓

Zheng, Chen, Lu and Zhou. CVPR 2019. Hardness-Aware Deep Metric Learning.
Ko and Gu. CVPR 2020. Embedding Expansion. Augmentation in Embedding Space for Deep Metric Learning.
Gu and Ko. 2020. Symmetrical Synthesis for Deep Metric Learning.
Gu, Ko and Kim. AAAI 2021. Proxy Synthesis: Learning with Synthetic Classes for Deep Metric Learning.
Kalantidis, Sariyildiz, Pion, Weinzaepfel and Larlus. NeurIPS 2020. Hard negative mixing for contrastive learning.
Lee, Zhu, Sohn, Li, Shin, and Lee. ICLR, 2021. I-Mix: A domain-agnostic strategy for contrastive representation learning.
Kim, Lee, Bae, and Yun. NeurIPS Workshops 2020. MixCo: Mix-up contrastive learning for visual representation.

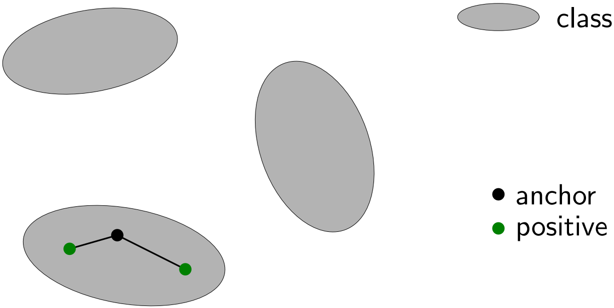
metrix (= metric mix)



- allow anchor to interact with positive examples (same class), negative examples (different class), and interpolated examples, which also have interpolated labels

Venkataraman *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

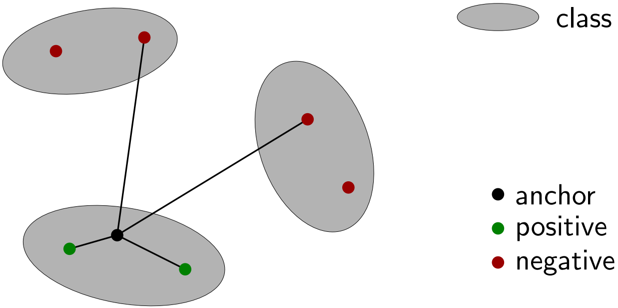
metrix (= metric mix)



- allow anchor to interact with positive examples (same class), negative examples (different class), and interpolated examples, which also have interpolated labels

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

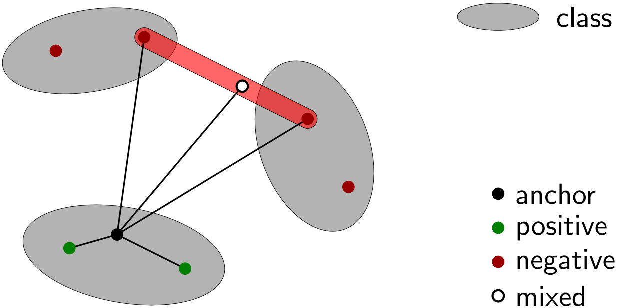
metrix (= metric mix)



- allow anchor to interact with positive examples (same class), negative examples (different class), and interpolated examples, which also have interpolated labels

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

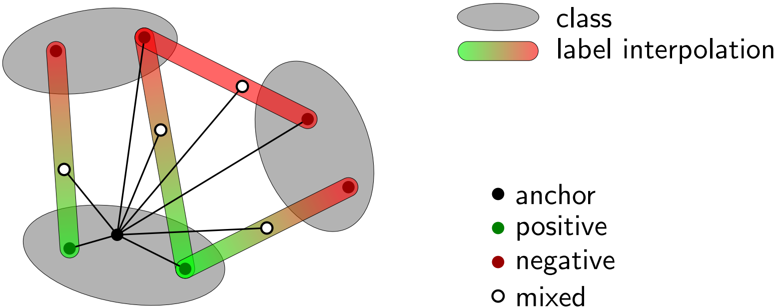
metrix (= metric mix)



- allow anchor to interact with positive examples (same class), negative examples (different class), and interpolated examples, which also have interpolated labels

Venkataraman *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

metrix (= metric mix)



- allow anchor to interact with positive examples (same class), negative examples (different class), and interpolated examples, which also have interpolated labels

Venkataraman *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

generic loss formulation

- contrastive loss $\ell_{\mathrm{C}}(a;\theta)$

$$\sum_{p\in P(a)} -s(a,p) + \sum_{n\in N(a)} [s(a,n)-m]_+$$

- multi-similarity loss $\ell_{\mathrm{MS}}(a;\theta)$

$$\frac{1}{\alpha}\log\left(1+\sum_{p\in P(a)} e^{-\alpha(s(a,p)-m)}\right)+\frac{1}{\beta}\log\left(1+\sum_{n\in N(a)} e^{\beta(s(a,n)-m)}\right)$$

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

generic loss formulation

- generic loss $\ell(a;\theta)$

$$\sigma^+\left(\sum_{p\in P(a)}\rho^+(s(a,p))\right)+\sigma^-\left(\sum_{n\in N(a)}\rho^-(s(a,n))\right)$$

- contrastive loss $\ell_{\mathrm{C}}(a;\theta)$

$$\sum_{p\in P(a)}-s(a,p)+\sum_{n\in N(a)}[s(a,n)-m]_+$$

- multi-similarity loss $\ell_{\mathrm{MS}}(a;\theta)$

$$\frac{1}{\alpha}\log\left(1+\sum_{p\in P(a)}e^{-\alpha(s(a,p)-m)}\right)+\frac{1}{\beta}\log\left(1+\sum_{n\in N(a)}e^{\beta(s(a,n)-m)}\right)$$

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

generic loss formulation

- generic loss $\ell(a;\theta)$

$$\sigma^{+}\left(\sum_{p\in P(a)} \rho^{+}(s(a,p))\right) + \sigma^{-}\left(\sum_{n\in N(a)} \rho^{-}(s(a,n))\right)$$

- contrastive loss $\ell_{\mathrm{C}}(a;\theta)$

$$\sum_{p\in P(a)} -s(a,p) + \sum_{n\in N(a)} [s(a,n)-m]_{+}$$

- multi-similarity loss $\ell_{\mathrm{MS}}(a;\theta)$

$$\frac{1}{\alpha}\log\left(1+\sum_{p\in P(a)} e^{-\alpha(s(a,p)-m)}\right) + \frac{1}{\beta}\log\left(1+\sum_{n\in N(a)} e^{\beta(s(a,n)-m)}\right)$$

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

generic loss formulation

- generic loss $\ell(a;\theta)$

$$\sigma^{+}\left(\sum_{p\in P(a)} \rho^{+}(s(a,p))\right) + \sigma^{-}\left(\sum_{n\in N(a)} \rho^{-}(s(a,n))\right)$$

- different loss functions in the generic formulation

Loss	Anchor	Pos/Neg	$\sigma^{+}(x)$	$\sigma^{-}(x)$	$\rho^{+}(x)$	$\rho^{-}(x)$
Contrastive	X	X	x	x	$-x$	$[x-m]_{+}$
Binomial deviance	X	X	$\log(1+x)$	$\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
Multi-similarity	X	X	$\frac{1}{\beta}\log(1+x)$	$\frac{1}{\gamma}\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
Proxy anchor	proxy	X	$\frac{1}{\beta}\log(1+x)$	$\frac{1}{\gamma}\log(1+x)$	$e^{-\beta(x-m)}$	$e^{\gamma(x-m)}$
NCA	X	X	$-\log(x)$	$\log(x)$	e^{x}	e^{x}
ProxyNCA	X	proxy	$-\log(x)$	$\log(x)$	e^{x}	e^{x}
ProxyNCA++	X	proxy	$-\log(x)$	$\log(x)$	$e^{x/T}$	$e^{x/T}$

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

mixing examples and labels

- generic loss $\ell(a;\theta)$

$$\sigma^+\left(\sum_{p\in P(a)} \rho^+(s(a,p))\right) + \sigma^-\left(\sum_{n\in N(a)} \rho^-(s(a,n))\right)$$

- defining $U(a) := \{(p,1) : p \in P(a)\} \cup \{(n,0) : n \in N(a)\}$,

$$\sigma^+\left(\sum_{(x,y)\in U(a)} y\rho^+(s(a,x))\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)$$

- defining $V(a) := \{(f_\lambda(x,x'), \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in U(a)^2\}$,

$$\sigma^+\left(\sum_{(v,y)\in V(a)} y\rho^+(s(a,v))\right) + \sigma^-\left(\sum_{(v,y)\in V(a)} (1-y)\rho^-(s(a,v))\right)$$

Venkataraman *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

mixing examples and labels

- generic loss $\ell(a;\theta)$

$$\sigma^+\left(\sum_{p\in P(a)} \rho^+(s(a,p))\right) + \sigma^-\left(\sum_{n\in N(a)} \rho^-(s(a,n))\right)$$

- defining $U(a) := \{(p,1) : p \in P(a)\} \cup \{(n,0) : n \in N(a)\}$,

$$\sigma^+\left(\sum_{(x,y)\in U(a)} y\rho^+(s(a,x))\right) + \sigma^-\left(\sum_{(x,y)\in U(a)} (1-y)\rho^-(s(a,x))\right)$$

- defining $V(a) := \{(f_\lambda(x,x'), \mathrm{mix}_\lambda(y,y')) : ((x,y),(x',y')) \in U(a)^2\}$,

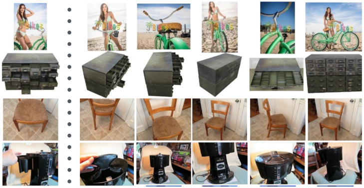
$$\sigma^+\left(\sum_{(v,y)\in V(a)} y\rho^+(s(a,v))\right) + \sigma^-\left(\sum_{(v,y)\in V(a)} (1-y)\rho^-(s(a,v))\right)$$

Venkataraman *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

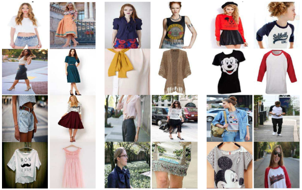
datasets

CUB

Cars



SOP



InShop

Wah, Branson, Welinder, Perona and Belongie. Caltech, 2011. The Caltech-UCSD Birds-200-2011 Dataset.
Krause, Stark, Deng and Fei-Fei. ICCVW 2013. 3D object representations for fine-grained categorization.
Song, Xiang, Jegelka and Savarese. CVPR 2016. Deep metric learning via lifted structured feature embedding.
Liu, Luo, Qiu, Wang and Tang. CVPR 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations.

R@k results with ResNet-50, $d = 512$

	CUB200		CARS196		SOP		IN-SHOP	
METHOD	R@1	R@2	R@1	R@2	R@1	R@10	R@1	R@10
Contrastive	64.7	75.9	81.6	88.2	74.9	87.0	86.4	94.7
+Metrix	67.4	77.9	85.1	91.1	77.5	89.1	89.1	95.7
	+2.7	+2.0	+3.5	+2.9	+2.6	+2.1	+2.7	+1.0
Multi-similarity	67.8	77.8	**87.8**	**92.7**	76.9	89.8	90.1	**97.6**
+Metrix	**71.4**	80.6	**89.6**	**94.2**	81.0	92.0	**92.2**	**98.5**
	+3.6	+2.8	+1.8	+1.5	+4.1	+2.2	+2.1	+0.9
Proxy anchor	**69.5**	79.3	87.6	92.3	79.1	90.8	90.0	97.4
+Metrix	71.0	**81.8**	89.1	93.6	**81.3**	91.7	91.9	98.2
	+1.3	+1.8	+1.4	+0.7	+2.2	+0.9	+1.9	+0.8
ProxyNCA++	69.1	**79.5**	86.6	92.1	**80.4**	**91.7**	**90.2**	**97.6**
+Metrix	70.4	80.6	88.5	93.4	**81.3**	**92.7**	91.9	98.1
	+1.3	+0.8	+1.9	+0.9	+0.6	+0.7	+1.5	+0.0
Gain over SOTA	+1.7	+1.8	+1.8	+1.3	+0.6	+0.0	+1.2	+0.4

Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

R@k results with ResNet-50, $d = 512$

	CUB200		CARS196		SOP		IN-SHOP	
METHOD	R@1	R@2	R@1	R@2	R@1	R@10	R@1	R@10
Contrastive	64.7	75.9	81.6	88.2	74.9	87.0	86.4	94.7
+Metrix	67.4	77.9	85.1	91.1	77.5	89.1	89.1	95.7
	+2.7	+2.0	+3.5	+2.9	+2.6	+2.1	+2.7	+1.0
Multi-similarity	67.8	77.8	**87.8**	**92.7**	76.9	89.8	90.1	**97.6**
+Metrix	**71.4**	80.6	**89.6**	**94.2**	81.0	92.0	**92.2**	**98.5**
	+3.6	+2.8	+1.8	+1.5	+4.1	+2.2	+2.1	+0.9
Proxy anchor	**69.5**	79.3	87.6	92.3	79.1	90.8	90.0	97.4
+Metrix	71.0	**81.8**	89.1	93.6	**81.3**	91.7	91.9	98.2
	+1.3	+1.8	+1.4	+0.7	+2.2	+0.9	+1.9	+0.8
ProxyNCA++	69.1	**79.5**	86.6	92.1	**80.4**	**91.7**	**90.2**	**97.6**
+Metrix	70.4	80.6	88.5	93.4	**81.3**	**92.7**	91.9	98.1
	+1.3	+0.8	+1.9	+0.9	+0.6	+0.7	+1.5	+0.0
Gain over SOTA	+1.7	+1.8	+1.8	+1.3	+0.6	+0.0	+1.2	+0.4

Kim, Kim, Cho and Kwak. CVPR 2020. Proxy anchor loss for deep metric learning.
Teh, DeVries and Taylor. ECCV 2020. ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis.
Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

R@k results with ResNet-50, $d = 512$

	CUB200		Cars196		SOP		In-Shop	
Method	R@1	R@2	R@1	R@2	R@1	R@10	R@1	R@10
Contrastive	64.7	75.9	81.6	88.2	74.9	87.0	86.4	94.7
+Metrix	67.4	77.9	85.1	91.1	77.5	89.1	89.1	95.7
	+2.7	+2.0	+3.5	+2.9	+2.6	+2.1	+2.7	+1.0
Multi-similarity	67.8	77.8	**87.8**	**92.7**	76.9	89.8	90.1	**97.6**
+Metrix	**71.4**	80.6	**89.6**	**94.2**	81.0	92.0	**92.2**	**98.5**
	+3.6	+2.8	+1.8	+1.5	+4.1	+2.2	+2.1	+0.9
Proxy anchor	**69.5**	79.3	87.6	92.3	79.1	90.8	90.0	97.4
+Metrix	71.0	**81.8**	89.1	93.6	**81.3**	91.7	91.9	98.2
	+1.3	+1.8	+1.4	+0.7	+2.2	+0.9	+1.9	+0.8
ProxyNCA++	69.1	**79.5**	86.6	92.1	**80.4**	**91.7**	**90.2**	**97.6**
+Metrix	70.4	80.6	88.5	93.4	**81.3**	**92.7**	91.9	98.1
	+1.3	+0.8	+1.9	+0.9	+0.6	+0.7	+1.5	+0.0
Gain over SOTA	+1.7	+1.8	+1.8	+1.3	+0.6	+0.0	+1.2	+0.4

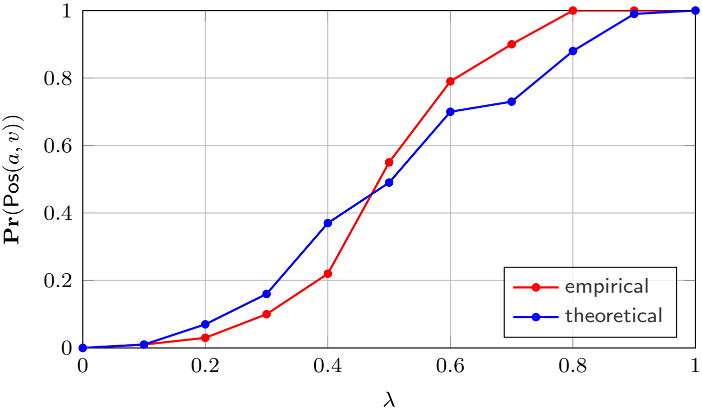
Hadsell, Chopra and LeCun. CVPR 2006. Dimensionality reduction by learning an invariant mapping.
Wang, Han, Huang, Dong, Scott. CVPR 2019. Multi-similarity loss with general pair weighting for deep metric learning.
Venkataramanan *et al.* 2021. It Takes Two to Tango: Mixup for Deep Metric Learning.

"positivity"

- $\text{Pos}(a, v)$: a mixed embedding v behaves as "positive" for anchor a: $\partial \ell(a; \theta) / \partial s(a, v) \leq 0$
- under certain assuptions, estimate the probability of $\text{Pos}(a, v)$ for a single mixed embedding v as a function of λ

"positivity"

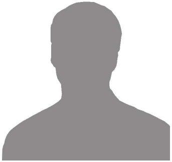
- Pos(a, v): a mixed embedding v behaves as "positive" for anchor a: $\partial \ell(a; \theta) / \partial s(a, v) \leq 0$
- under certain assuptions, estimate the probability of Pos(a, v) for a single mixed embedding v as a function of λ



attention

global-local, spatial-channel attention for image retrieval

[WACV 2022]



Chull Hwan Song

Hye Joo Han

Yannis Avrithis

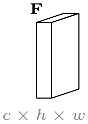
paper

`https://arxiv.org/abs/2107.08000`

code

by WACV (January)

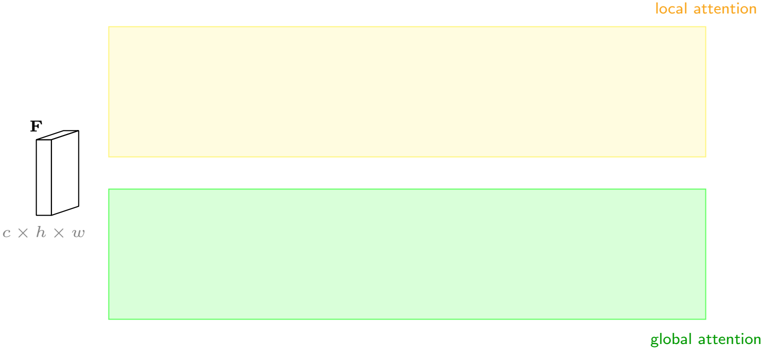
global-local attention module (GLAM)



- input feature tensor: c feature maps (channels), $h \times w$ spatial resolution

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

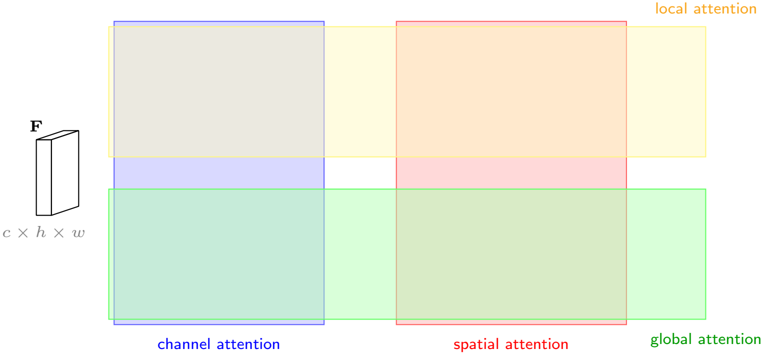
global-local attention module (GLAM)



- local (1st order) attention: elements of the feature tensor (channels / spatial locations) weighted independently, by pooling or learning
- global (2nd order) attention: pairwise interaction between elements of the tensor

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

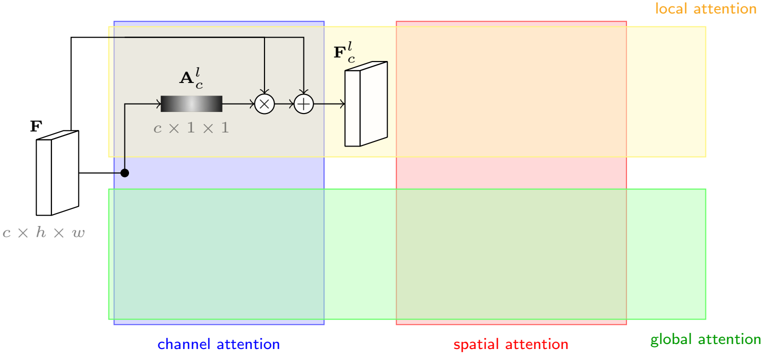
global-local attention module (GLAM)



- channel attention: channels weighted independently or interact pairwise
- spatial attention: spatial locations weighted independently or interact pairwise

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

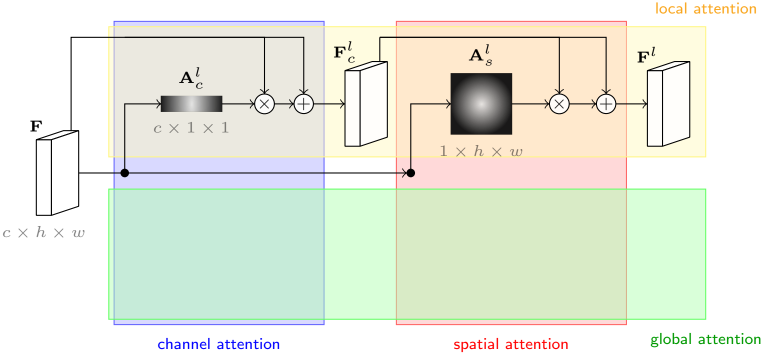
global-local attention module (GLAM)



- local channel attention: pooling over locations yields $c \times 1 \times 1$ attention map
-

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

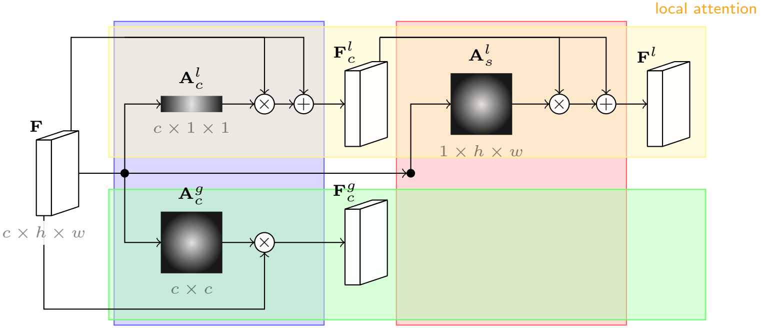
global-local attention module (GLAM)



- local channel attention: pooling over locations yields $c \times 1 \times 1$ attention map
- local spatial attention: pooling over channels yields $1 \times h \times w$ attention map

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

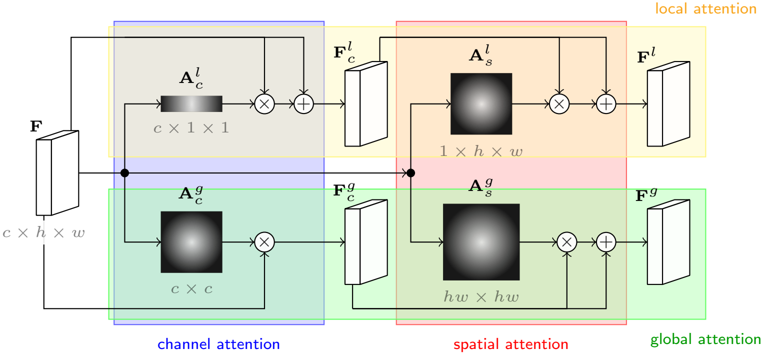
global-local attention module (GLAM)



- global channel attention: pairwise interaction of channels yields $c \times c$ attention map
-

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

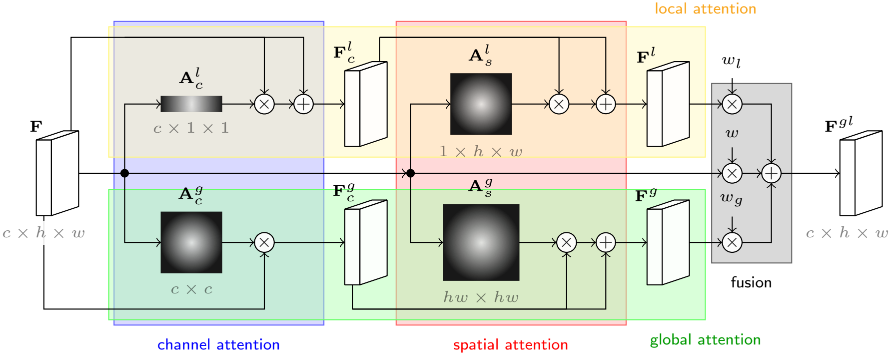
global-local attention module (GLAM)



- global channel attention: pairwise interaction of channels yields $c \times c$ attention map
- global spatial attention: pairwise interaction of locations yields $hw \times hw$ attention map

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

global-local attention module (GLAM)



- fusion: local and global attention streams fused with original feature tensor

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

image retrieval study

- ResNet101 backbone, GeM pooling
- global descriptor only, $d = 512$
- train by Arcface loss on Google Landmarks v2 clean (1.5M images)
- mini-batch examples with similar aspect ratios resized jointly
- at inference, multi-resolution representation to queries and database
- test on Revisited Oxford ($\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Par)
- ablate local/global, channel/spatial attention components

Radenović, Iscen, Tolias, Avrithis and Chum. CVPR 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking.
Yokoo, Ozaki, Simo-Serra and Iizuka. CVPRW 2020. Two-stage Discriminative Re-ranking for Large-scale Landmark Retrieval.
Weyand, Araujo, Cao and Sim. CVPR 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval.
Deng, Guo, Xue and Zafeiriou. CVPR 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition.
Radenović, Tolias and Chum. TPAMI, 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation.
Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

image retrieval study

- ResNet101 backbone, GeM pooling
- global descriptor only, $d = 512$
- train by Arcface loss on Google Landmarks v2 clean (1.5M images)
- mini-batch examples with similar aspect ratios resized jointly
- at inference, multi-resolution representation to queries and database
- test on Revisited Oxford ($\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Par)
- ablate local/global, channel/spatial attention components

Radenović, Iscen, Tolias, Avrithis and Chum. CVPR 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking.
Yokoo, Ozaki, Simo-Serra and Iizuka. CVPRW 2020. Two-stage Discriminative Re-ranking for Large-scale Landmark Retrieval.
Weyand, Araujo, Cao and Sim. CVPR 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval.
Deng, Guo, Xue and Zafeiriou. CVPR 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition.
Radenović, Tolias and Chum. TPAMI, 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation.
Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

image retrieval study

- ResNet101 backbone, GeM pooling
- global descriptor only, $d = 512$
- train by Arcface loss on Google Landmarks v2 clean (1.5M images)
- mini-batch examples with similar aspect ratios resized jointly
- at inference, multi-resolution representation to queries and database
- test on Revisited Oxford ($\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Par)
- ablate local/global, channel/spatial attention components

Radenović, Iscen, Tolias, Avrithis and Chum. CVPR 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking.
Yokoo, Ozaki, Simo-Serra and Iizuka. CVPRW 2020. Two-stage Discriminative Re-ranking for Large-scale Landmark Retrieval.
Weyand, Araujo, Cao and Sim. CVPR 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval.
Deng, Guo, Xue and Zafeiriou. CVPR 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition.
Radenović, Tolias and Chum. TPAMI, 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation.
Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

ablation

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- channel/spatial attention: may be harmful when used alone, but complementary and surprisingly beneficial when used together
- local/global attention: clearly complementary, their gain nearly additive

ablation

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- channel/spatial attention: may be harmful when used alone, but complementary and surprisingly beneficial when used together
- local/global attention: clearly complementary, their gain nearly additive

ablation

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- channel/spatial attention: may be harmful when used alone, but complementary and surprisingly beneficial when used together
- local/global attention: clearly complementary, their gain nearly additive

ablation

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- channel/spatial attention: may be harmful when used alone, but complementary and surprisingly beneficial when used together
- local/global attention: clearly complementary, their gain nearly additive

ablation

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- channel/spatial attention: may be harmful when used alone, but complementary and surprisingly beneficial when used together
- local/global attention: clearly complementary, their gain nearly additive

thank you!

more

https://avrithis.net