Matching Feature Sets for Few-Shot Image Classification

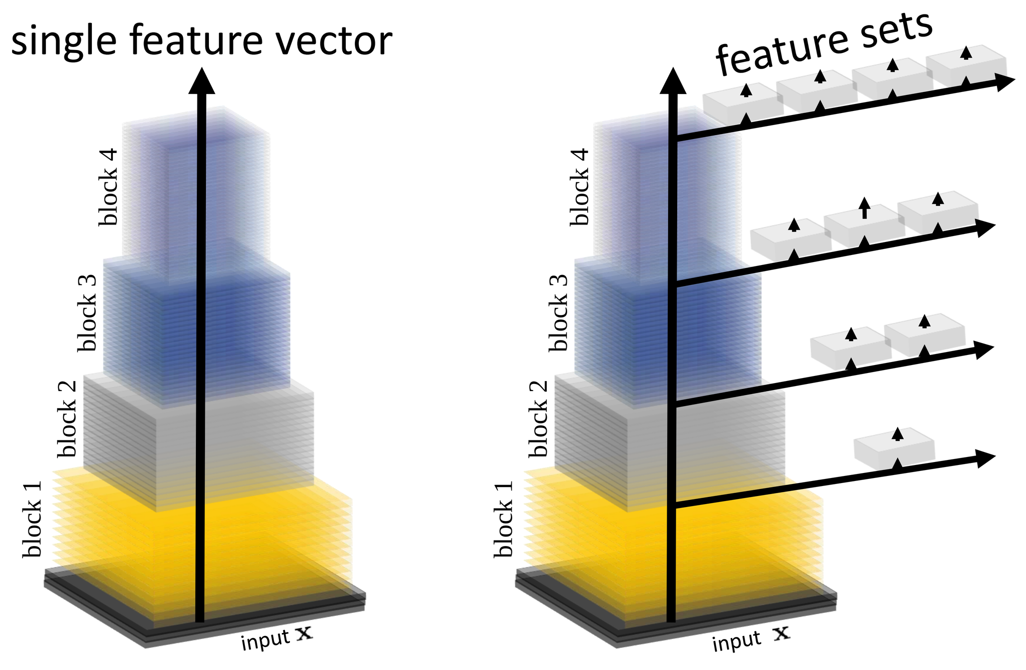
Arman Afrasiyabi⋆•, Hugo Larochelle⋄†•, Jean-François Lalonde⋆, Christian Gagné⋆†•

⋆Université Laval, ⋄Google Brain, †Canada CIFAR AI Chair, •Mila

https://lvsn.github.io/SetFeat/

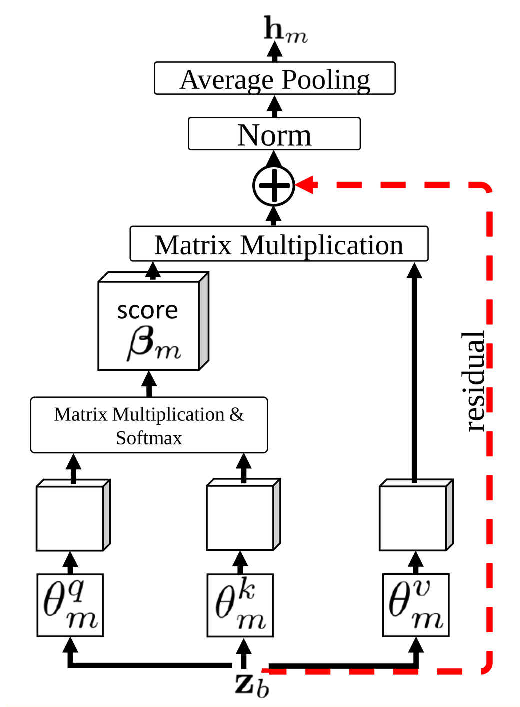
SetFeat

- it is common practice to extract a single feature vector per input image
- we propose set-feature extractor (SetFeat) to represent images as *sets* of feature vectors



a) common practice b) ours

- we take inspiration from Feature Pyramid Networks to learn a richer feature space
- SetFeat embeds shallow self-attention mappers in existing architecture



Mapper $g(\cdot|\theta^g_m)$

- for attention map, we first compute

$$\beta_m = \mathrm{Softmax}\left(q(\mathbf{z}_{b_m}|\theta^q_m)k(\mathbf{z}_{b_m}|\theta^k_m)^\top/\sqrt{d_k}\right) ,$$

where $\beta_m \in \mathbb{R}^{P\times P}$

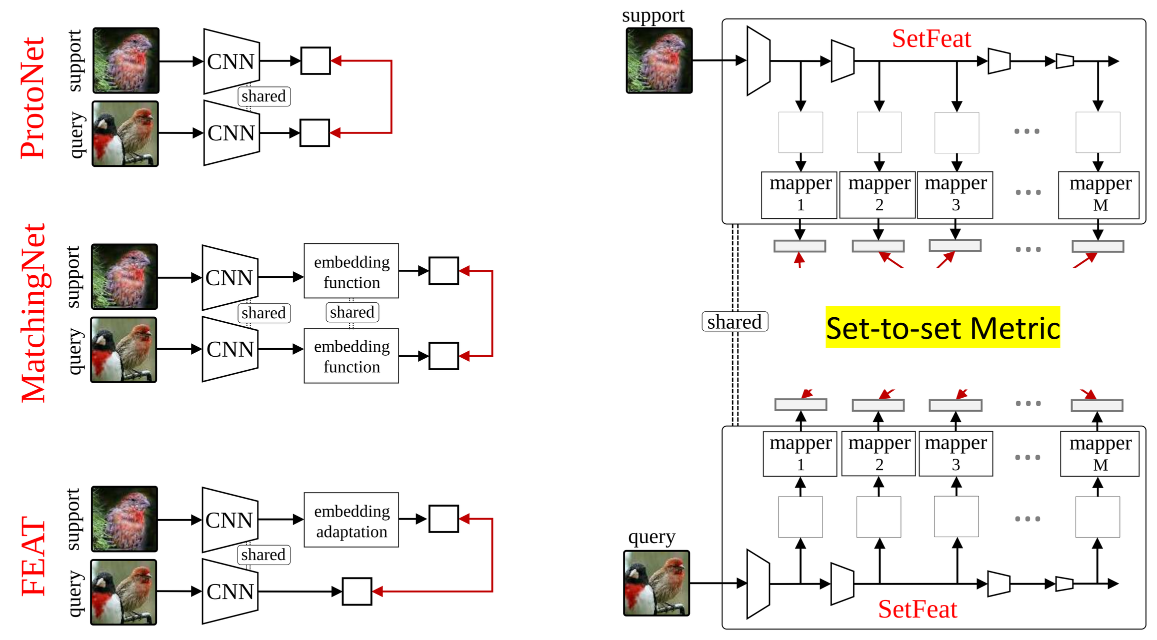
- then, we compute the attention over β_m

$$\mathbf{a}_m = \beta_m\, v(\mathbf{z}_{b_m}|\theta^v_m) ,$$

where $\mathbf{a}_m \in \mathbb{R}^{P\times D^a}$ consists of P patches

Set-to-set metrics

- SetFeat first extracts sets of features
- then, we need a set-to-set metric to compare the feature set of the query with the feature sets corresponding to each instance of the support set of each class



a) Three baseline methods b) Set matching (ours)

- **Match-sum** aggregates the distance between matching mappers

$$d_{\mathrm{ms}}(\mathbf{x}_q, \mathcal{S}^n) = \sum_{i=1}^{M} d\left(\mathbf{h}_i(\mathbf{x}_q), \bar{\mathbf{h}}_i(\mathcal{S}^n)\right) .$$

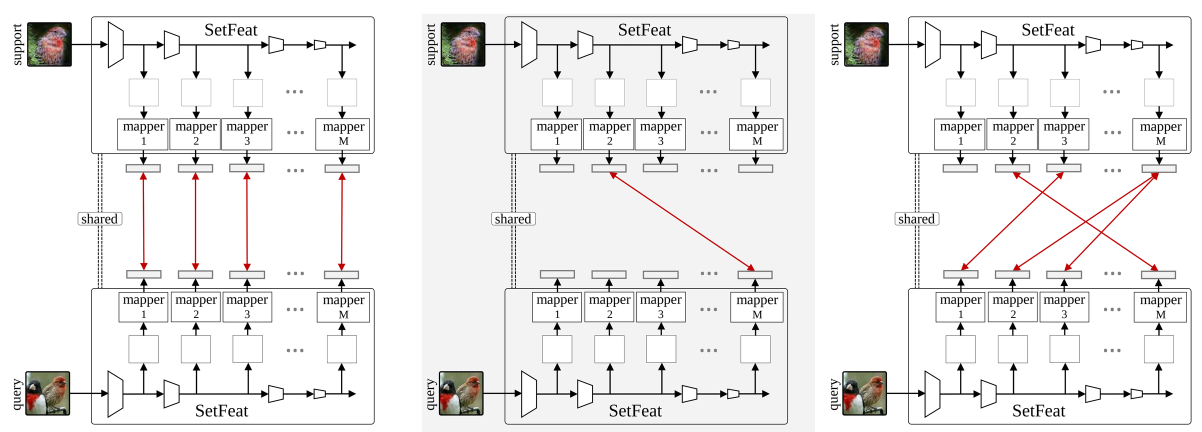
- **Min-min** uses the minimum distance across all possible pairs of elements

$$d_{\mathrm{mm}}(\mathbf{x}_q, \mathcal{S}^n) = \min_{i=1}^{M} \min_{j=1}^{M} d\left(\mathbf{h}_i(\mathbf{x}_q), \bar{\mathbf{h}}_j(\mathcal{S}^n)\right) .$$

- **Sum-min** aggregates with a sum the minimum distances between the mappers

$$d_{\mathrm{sm}}(\mathbf{x}_q, \mathcal{S}^n) = \sum_{i=1}^{M} \min_{j=1}^{M} d\left(\mathbf{h}_i(\mathbf{x}_q), \bar{\mathbf{h}}_j(\mathcal{S}^n)\right) .$$

- illustration of our set-to-set metric



(a) match-sum (b) min-min (c) sum-min

Evaluations

- MiniImageNet evaluations of SetFeat4 results in +**2.03**% improvement in 1-shot

Table 1. Evaluation on miniImageNet in 5-way. Bold/blue is best/second, and ± is the 95% confidence intervals in 600 episodes.

	Method	Backbone	1-shot	5-shot
	ProtoNet [50]	Conv4-64	49.42±0.78	68.20±0.66
	MAML [18]	Conv4-64	48.07±1.75	63.15±0.91
	RelationNet [53]	Conv4-64	50.44±0.82	65.32±0.70
	Baseline++ [8]	Conv4-64	48.24±0.75	66.43±0.63
	IMP [3]	Conv4-64	49.60±0.80	68.10±0.80
	MemoryNet [7]	Conv4-64	53.37±0.48	66.97±0.35
	Neg-Margin [35]	Conv4-64	52.84±0.76	70.41±0.66
	MixtFSL [2]	Conv4-64	52.82±0.63	70.67±0.57
	FEAT [68]	Conv4-64	55.15±0.20	71.61±0.16
	MELR [16]	Conv4-64	55.35±0.43	72.27±0.35
	BOIL [39]	Conv4-64	49.61±0.16	66.45±0.37
Ours	Match-sum	SF4-64	55.74±0.65	72.18±0.70
Ours	Min-min	SF4-64	56.22±0.89	72.70±0.65
Ours	Sum-min	SF4-64	**57.18**±0.89	**73.67**±0.71

- TieredImageNet evaluation of SetFeat12 results in +**1.42**% improvement in 1-shot

Table 2. TieredImageNet evaluation. Bold/red is best/second best, and ± indicates the 95% conf. intervals over 600 episodes of 5-way.

	Method	Backbone	1-shot	5-shot
	OptNet [31]	ResNet12	65.99±0.72	81.56±0.53
	MTL [52]	ResNet12	65.62±1.80	80.61±0.90
	DNS [48]	ResNet12	66.22±0.75	82.79±0.48
	Simple [55]	ResNet12	69.74±0.72	84.41±0.55
	TapNet [70]	ResNet12	63.08±0.15	80.26±0.12
	ProtoNet† [50]	ResNet12	68.23± 0.23	84.03±0.16
	FEAT [68]	ResNet12	70.80±0.23	84.79±0.16
	MixtFSL [2]	ResNet12	70.97±1.03	86.16±0.67
	Distill [55]	ResNet12	71.52±0.69	86.03±0.49
	DeepEMD [74]	ResNet12	71.16±0.87	86.03±0.58
	DMF [66]	ResNet12	71.89±0.52	85.96±0.35
	MELR [16]	ResNet12	72.14±0.51	87.01±0.35
	Distill [45]	ResNet12	72.21±0.90	87.08±0.58
Ours	Match-sum	SF12	71.22±0.86	85.43±0.55
Ours	Min-min	SF12	71.75±0.90	86.40±0.56
Ours	Sum-min	SF12	**73.63**±0.88	**87.59**±0.57

†taken from [31]; Mappers dimension: SF12 $\in \mathbb{R}^{512}$

- CUB evaluation of SetFeat4 results in +**1.83**% improvement in 1-shot

Table 3. Fine-grained evaluation using CUB in 5-way. ± is the 95% confidence intervals on 600 episodes(‡taken from [54]).

	Method	Backbone	1-shot	5-shot
	MatchingNet [59]	Conv4-64	61.16±0.89	72.86±0.70
	ProtoNet [50]	Conv4-64	64.42±0.48	81.82±0.35
	MAML [17]	Conv4-64	55.92±0.95	72.09±0.76
	RelationNet [53]	Conv4-64	62.45±0.98	76.11±0.69
	FEAT [68]	Conv4-64	68.87±0.22	82.90±0.15
	MELR [16]	Conv4-64	70.26±0.50	85.01±0.32
Ours	Match-sum	SF4-64	67.35±0.93	83.82±0.61
Ours	Min-min	SF4-64	70.15±0.93	84.94±0.64
Ours	Sum-min	SF4-64	**72.09**±0.92	**87.05**±0.58

Ablations

- Mapper configurations. different ways of embedding ten mappers throughout the backbone

Table 4. Ablation of different mapper-level combinations using miniImageNet. The results are validation accuracy with min-sum.

Mappers	SetFeat4-64 1-shot	SetFeat4-64 5-shot	SetFeat4-512 1-shot	SetFeat4-512 5-shot
ProtoNet*	53.51	71.57	–	–
0-0-0-1	53.55	71.51	–	–
1-2-3-4 (concat)	53.56	71.82	–	–
1-1-1-1	51.11	69.41	53.57	71.60
0-0-0-10	52.90	69.49	55.36	71.59
2-2-3-3	54.73	71.98	56.29	74.74
1-2-3-4	54.71	71.35	58.74	75.30

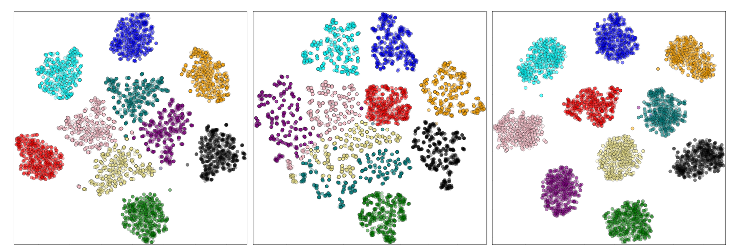
* with Conv4-512

- Top-m analysis. the results improve as we move towards sum-min, which uses all of the mappers

Table 6. Ablation of top-m mapper in the min-sum metric using SetFeat4 and SetFeat12* on CUB. The results are validation set.

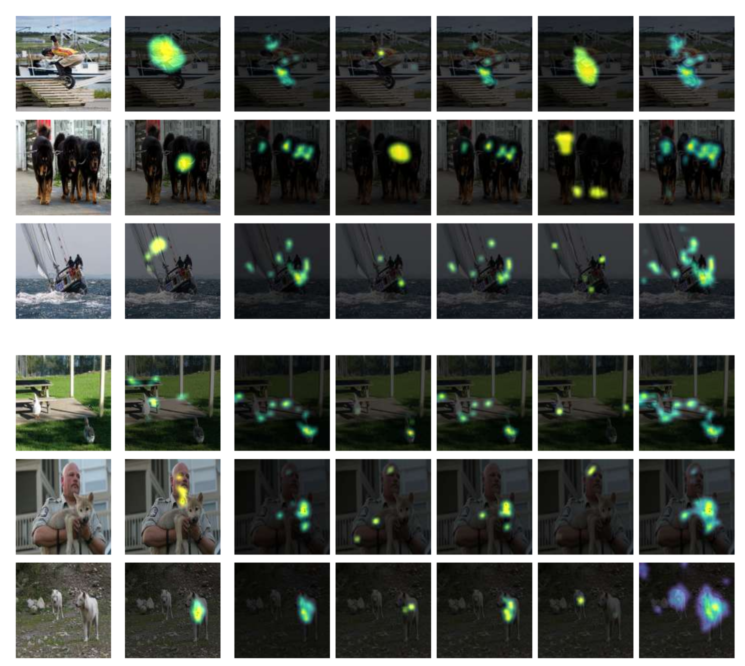
Method	SetFeat4 1-shot	SetFeat4 5-shot	SetFeat12* 1-shot	SetFeat12* 5-shot
top-1 (min-min)	70.15	84.94	78.51	89.73
top-2	70.84	85.30	77.92	89.87
top-4	70.34	85.95	78.37	89.78
top-8	71.47	86.88	79.56	90.03
top-10 (sum-min)	**72.09**	**87.05**	**79.60**	**90.48**

- t-SNE visualization. the distributions of mapper embeddings are generally disjoint



(a) miniImageNet (b) CUB (c) tieredImageNet

- Visualizing mappers saliency. our approach devotes attention to many more parts of the images



(a) input (b) baseline (c) ours

Acknowledgements. This project was supported by NSERC-Canada and CIFAR Chair.