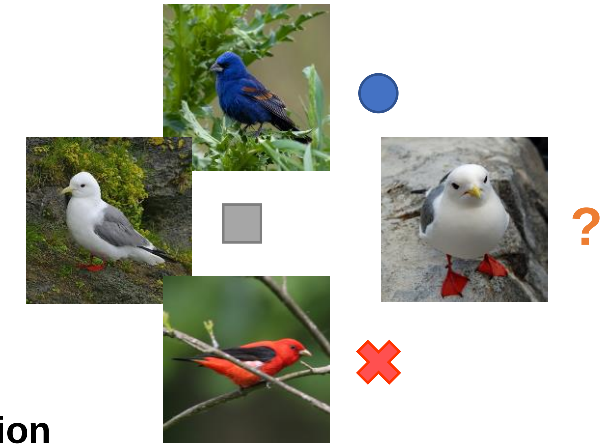
Attributes-Guided and Pure-Visual Attention Alignment for Few-Shot Recognition

Siteng Huang, Min Zhang, Yachen Kang, Donglin Wang*

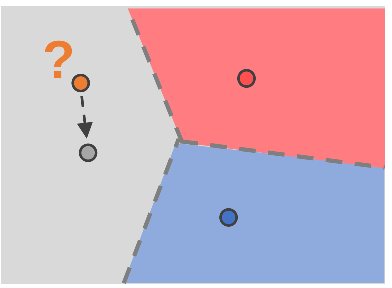
{huangsiteng, wangdonglin}@westlake.edu.cn

Background



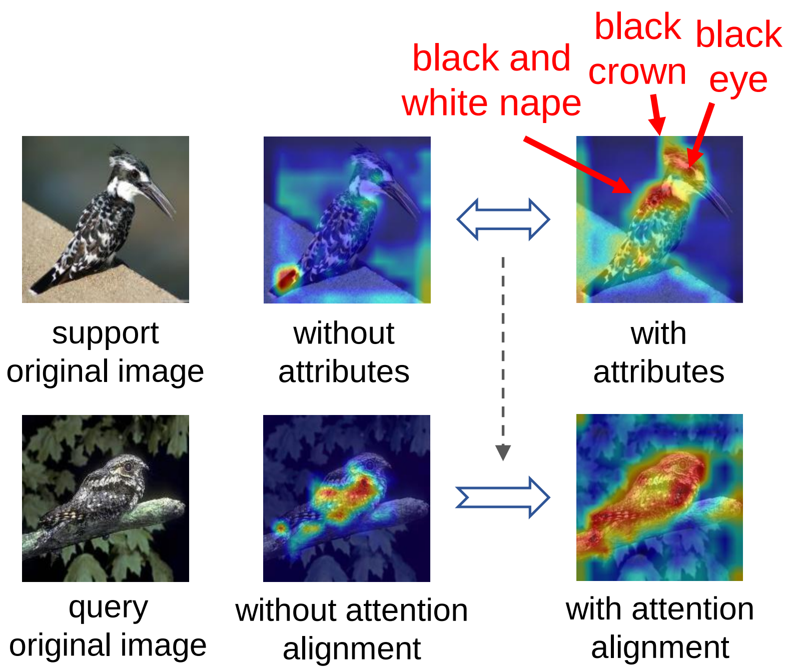
Few-shot recognition: recognize novel categories with very few labeled examples in each class.

Poor generalization

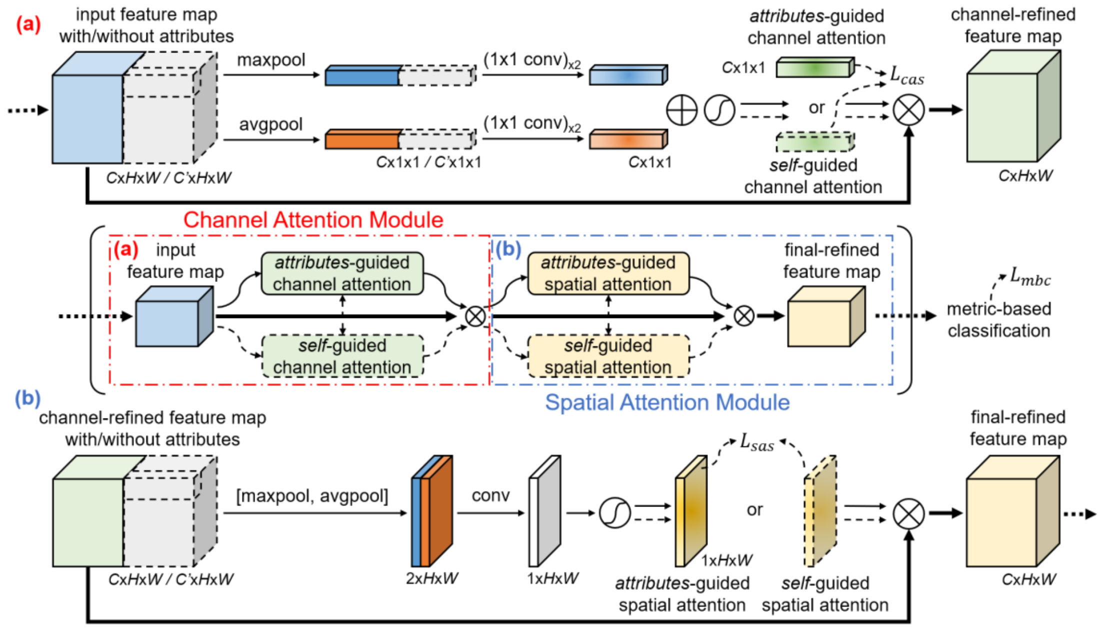


Metric-based meta-learning: learn a generalizable embedding model to transform all samples into a common metric space, where simple nearest-neighbor classifiers can be executed.

In this paper, we propose a novel **attributes-guided attention module (AGAM)** to utilize human-annotated attributes as auxiliary semantics and learn more discriminative features.



Approach



1. We design two parallel branches – attributes-guided branch for samples with attributes, and self-guided branch for samples without attributes. Discriminability of features is improved with attributes-guided or self-guided channel and spatial attention.

2. Similar feature selection processes are proposed for both support and query samples, so features extracted by both visual contents and attributes share the same space with pure-visual features.

3. We propose an attention alignment mechanism between two branches, promoting the self-guided branch to focus on more important features even without attributes.

Experimental Results

Extensive experiments show that our light-weight module can significantly improve metric-based approaches to achieve SOTA. More details can be found in

- **Project Page**: https://kyonhuang.top/publication/attributes-guided-attention-module
- **Code**: https://github.com/bighuang624/AGAM

Method	CUB		SUN	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016), *paper*	61.16 ± 0.89	72.86 ± 0.70	-	-
MatchingNet (Vinyals et al. 2016), *our implementation*	62.82 ± 0.36	73.22 ± 0.23	55.72 ± 0.40	76.59 ± 0.21
MatchingNet (Vinyals et al. 2016) **with AGAM**	**71.58 ± 0.30** *+8.76*	**75.46 ± 0.28** *+2.24*	**64.95 ± 0.35** *+9.23*	**79.06 ± 0.19** *+2.47*
ProtoNet (Snell, Swersky, and Zemel 2017), *paper*	51.31 ± 0.91	70.77 ± 0.69	-	-
ProtoNet (Snell, Swersky, and Zemel 2017), *our implementation*	53.01 ± 0.34	71.91 ± 0.22	57.76 ± 0.29	79.27 ± 0.19
ProtoNet (Snell, Swersky, and Zemel 2017) **with AGAM**	**75.87 ± 0.29** *+22.86*	**81.66 ± 0.25** *+9.75*	**65.15 ± 0.31** *+7.39*	**80.08 ± 0.21** *+0.81*
RelationNet (Sung et al. 2018), *paper*	62.45 ± 0.98	76.11 ± 0.69	-	-
RelationNet (Sung et al. 2018), *our implementation*	58.62 ± 0.37	78.98 ± 0.24	49.58 ± 0.35	76.21 ± 0.19
RelationNet (Sung et al. 2018) **with AGAM**	**66.98 ± 0.31** *+8.36*	**80.33 ± 0.40** *+1.35*	**59.05 ± 0.32** *+9.47*	**77.52 ± 0.18** *+1.31*

Table 1: Average accuracy (%) comparison with 95% confidence intervals before and after incorporating AGAM into existing methods using a Conv-4 backbone. Best results are displayed in **boldface**, and improvements are displayed in *italics*.

Method	Backbone	Test Accuracy	
		5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016)	Conv-4	61.16 ± 0.89	72.86 ± 0.70
ProtoNet (Snell, Swersky, and Zemel 2017)	Conv-4	51.31 ± 0.91	70.77 ± 0.69
RelationNet (Sung et al. 2018)	Conv-4	62.45 ± 0.98	76.11 ± 0.69
MACO (Hilliard et al. 2018)	Conv-4	60.76	74.96
MAML (Finn, Abbeel, and Levine 2017)	Conv-4	55.92 ± 0.95	72.09 ± 0.76
Baseline (Chen et al. 2019a)	Conv-4	47.12 ± 0.74	64.16 ± 0.71
Baseline++ (Chen et al. 2019a)	Conv-4	60.53 ± 0.83	79.34 ± 0.61
Comp. (Tokmakov, Wang, and Hebert 2019) *	ResNet-10	53.6	74.6
AM3 (Xing et al. 2019) † *	Conv-4	73.78 ± 0.28	81.39 ± 0.26
AGAM (OURS) *	Conv-4	**75.87 ± 0.29**	**81.66 ± 0.25**
MatchingNet (Vinyals et al. 2016) †	ResNet-12	60.96 ± 0.35	77.31 ± 0.25
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	68.8	76.4
RelationNet (Sung et al. 2018) †	ResNet-12	60.21 ± 0.35	80.18 ± 0.25
TADAM (Oreshkin, López, and Lacoste 2018)	ResNet-12	69.2	78.6
FEAT (Ye et al. 2020)	ResNet-12	68.87 ± 0.22	82.90 ± 0.15
MAML (Finn, Abbeel, and Levine 2017)	ResNet-18	69.96 ± 1.01	82.70 ± 0.65
Baseline (Chen et al. 2019a)	ResNet-18	65.51 ± 0.87	82.85 ± 0.55
Baseline++ (Chen et al. 2019a)	ResNet-18	67.02 ± 0.90	83.58 ± 0.54
Delta-encoder (Bengio et al. 2018)	ResNet-18	69.8	82.6
Dist. ensemble (Dvornik, Mairal, and Schmid 2019)	ResNet-18	68.7	83.5
SimpleShot (Wang et al. 2019)	ResNet-18	70.28	86.37
AM3 (Xing et al. 2019) *	ResNet-12	73.6	79.9
Multiple-Semantics (Schwartz et al. 2019) * ° •	DenseNet-121	76.1	82.9
Dual TriNet (Chen et al. 2019b) * °	ResNet-18	69.61 ± 0.46	84.10 ± 0.35
AGAM (OURS) *	ResNet-12	**79.58 ± 0.25**	**87.17 ± 0.23**

Table 2: Average accuracy (%) comparison to state-of-the-arts with 95% confidence intervals on the CUB dataset. † denotes that it is our implementation. * denotes that it uses auxiliary attributes. ° denotes that it uses auxiliary label embeddings. • denotes that it uses auxiliary descriptions of the categories. Best results are displayed in **boldface**.

Method	Backbone	Test Accuracy	
		5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016) †	Conv-4	55.72 ± 0.40	76.59 ± 0.21
ProtoNet (Snell, Swersky, and Zemel 2017) †	Conv-4	57.76 ± 0.29	79.27 ± 0.19
RelationNet (Sung et al. 2018) †	Conv-4	49.58 ± 0.35	76.21 ± 0.19
Comp. (Tokmakov, Wang, and Hebert 2019) *	ResNet-10	45.9	67.1
AM3 (Xing et al. 2019) † *	Conv-4	62.79 ± 0.32	79.69 ± 0.23
AGAM (OURS) *	Conv-4	**65.15 ± 0.31**	**80.08 ± 0.21**

Table 3: Average accuracy (%) comparison to state-of-the-arts with 95% confidence intervals on the SUN dataset. † denotes that it is our implementation. * denotes that it uses auxiliary attributes. Best results are displayed in **boldface**.