Zhiyu Xue

Undergraduate Student
Data Intelligence Group, UESTC

电子科技大学
University of Electronic Science and Technology of China

Name: Zhiyu Xue (Chris)

College: University of Electronic Science and Technology of China

Major: Data Science and Big Data Technology (School of Computer Science and Engineering)

GPA: 3.78/4.0

TOEFL: 86/100 (Taken at Sept. 2019)

Supervisors: Lixin Duan, Wen Li

Research: few-shot learning, interpretability, image captioning

CONTENTS

Research Experiences

Relative Position and Map Networks in Few-shot Learning for Image Classification

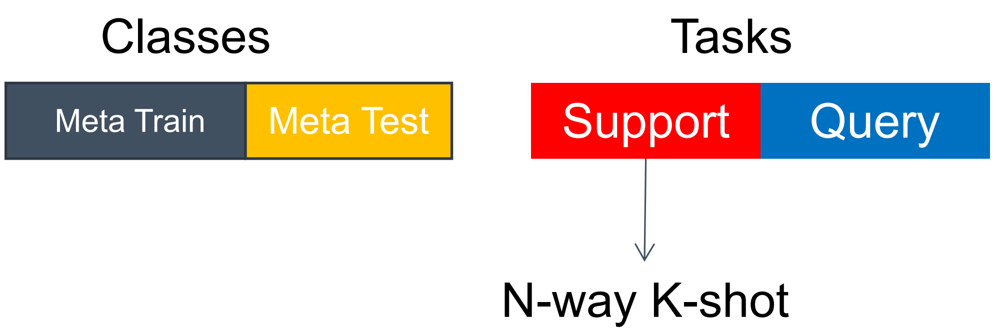
Zhiyu Xue, Zhenshan Xie, Zheng Xing, Lixin Duan
UESTC

Accepted by CVPR 2020 VL3 Workshop
Codes: *https://github.com/chrisyxue/RMN-RPN-for-FSL.*

Few-shot Learning

- Normal Training:

- Few-shot Training (Meta Training):



Baseline

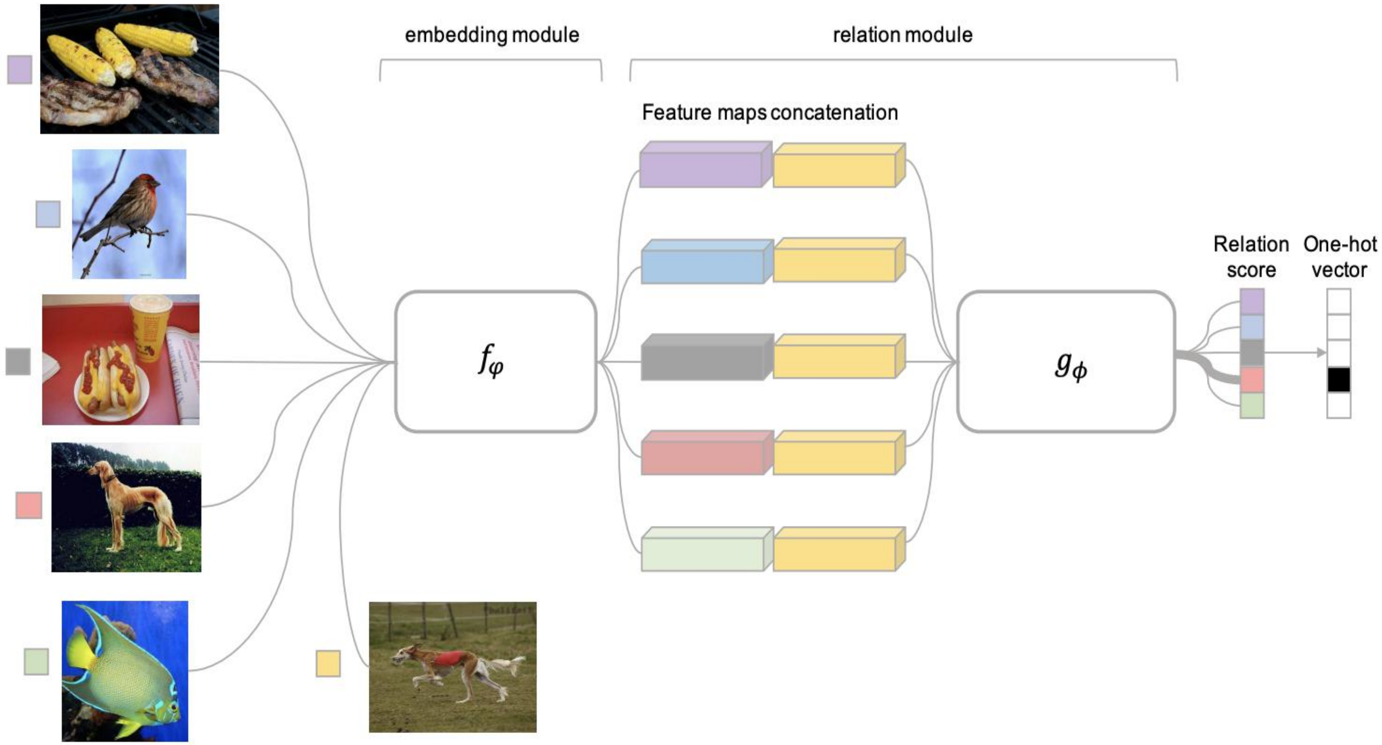
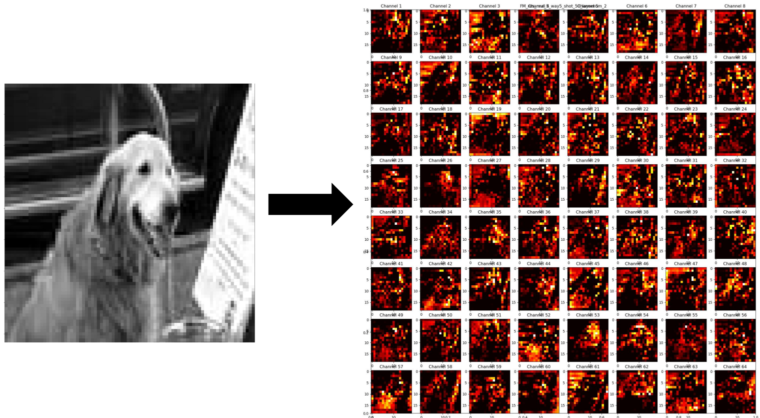


Figure 1: Relation Network architecture for a 5-way 1-shot problem with one query example.

Sung, Flood, et al. "Learning to compare: Relation network for few-shot learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

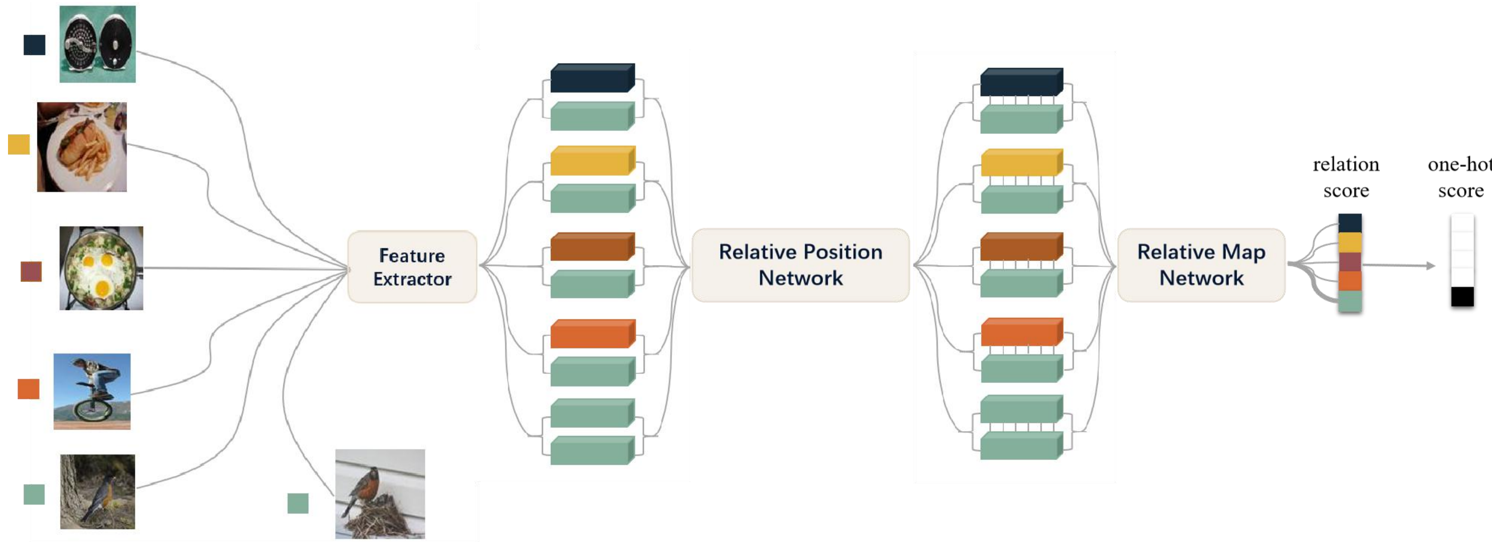
Motivation

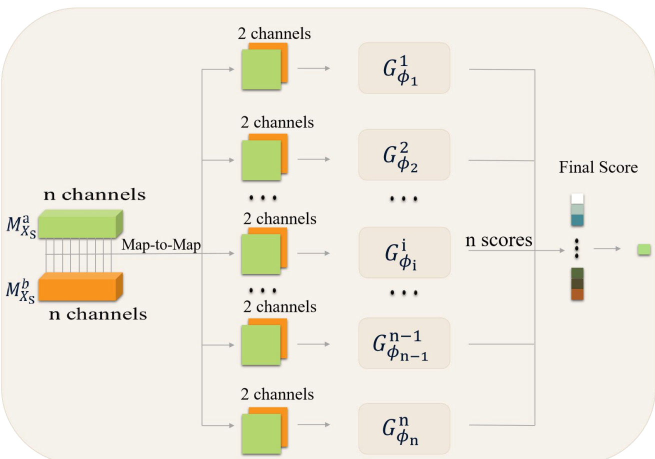


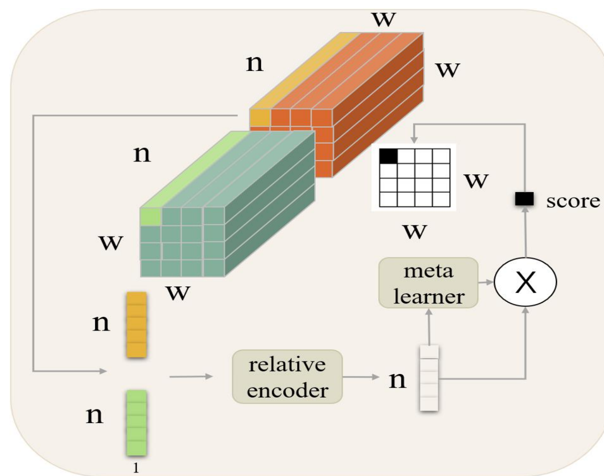
RMN:different channels have different descriptions

RPN: the importance of each position is different

Architecture







$$P_{S,Q} = Sig(\sum_{i=1}^{n} w_i G^i_{\phi_i}(M^i_{x_S}, M^i_{x_Q}))$$

$$V^{s,q}_{i,j} = H([v^S_{i,j}, v^Q_{i,j}])$$

$$Att_{i,j} = w^T V^{s,q}_{i,j}$$

$$w = W_2 \cdot \sigma(W_1 \cdot V^{s,q}_{i,j})$$

$$M_{x^Q} := M_{x^Q} + Att \otimes M_{x^Q}$$

Experiments

Table 1. Mean accuracies (%) of different methods on the MiniImageNet dataset. Results are obtained over 600 test episodes with 95% confidence intervals.

Model	MiniImageNet (5-way)	
	1-shot	5-shot
MATCHING NETS [21]	43.56±0.84	55.31±0.73
META LSTM [15]	43.44±0.77	60.60±0.71
MAML [3]	48.70±1.84	63.11±0.92
PROTOTYPICAL NETS [19]	49.42±0.78	68.20±0.66
META SGD [12]	50.47±1.87	64.03±0.94
RN [20]	50.44±0.82	65.32±0.70
GNN [17]	50.33±0.36	66.41±0.63
PABN [6]	51.87	65.37
TPN [13]	52.78±0.27	66.59±0.28
EGNN(No Trans) [8]	-	66.85
R2-D2 [2]	51.80±0.20	68.4±0.20
Ours(Conv4)	51.72±0.67	67.80±0.30
Ours(Our backbone)	**53.35± 0.77**	**69.35± 0.61**

Table 2. Mean accuracies (%) of different methods on the CIFAR-FS dataset. Results are obtained over 600 test episodes with 95% confidence intervals.

Model	CIFAR-FS (5-way)	
	1-shot	5-shot
MAML [3]	58.9±1.9	71.5±1.0
PROTOTYPICAL NETS [19]	55.5±0.7	72.0±0.6
RN [20]	55.0±1.0	69.3±0.8
GNN [17]	61.9	75.3
R2-D2 [2]	62.3±0.2	77.4±0.2
Ours	**61.43**	**76.16**

Table 3. Ablation study w.r.t. average accuracies (%) over 600 test episodes with 95% confidence intervals MiniImageNet in task 5-way K-shot about ablation study, where $K = 1, 3, 5, 7$ and 10.

Ave Acc	5-1	5-3	5-5	5-7	5-10
RN [20]	50.44	60.63	65.32	67.73	69.81
RPN	52.43	62.96	67.03	69.51	72.01
RMN	50.54	63.12	68.28	70.49	72.12
Ours	**53.35**	**63.94**	**69.35**	**70.87**	**73.17**

Region Comparison Network for Interpretable Few-shot Image Classification

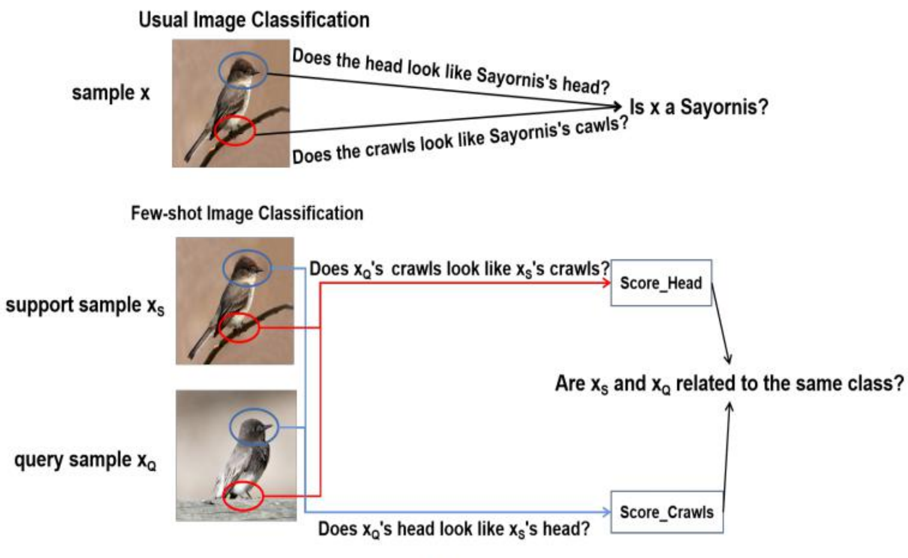
Zhiyu Xue, Wen Li, Lixin Duan, Lin Chen, Jiebo Luo
UESTC, Futurewei, UR

Finished in May 2020

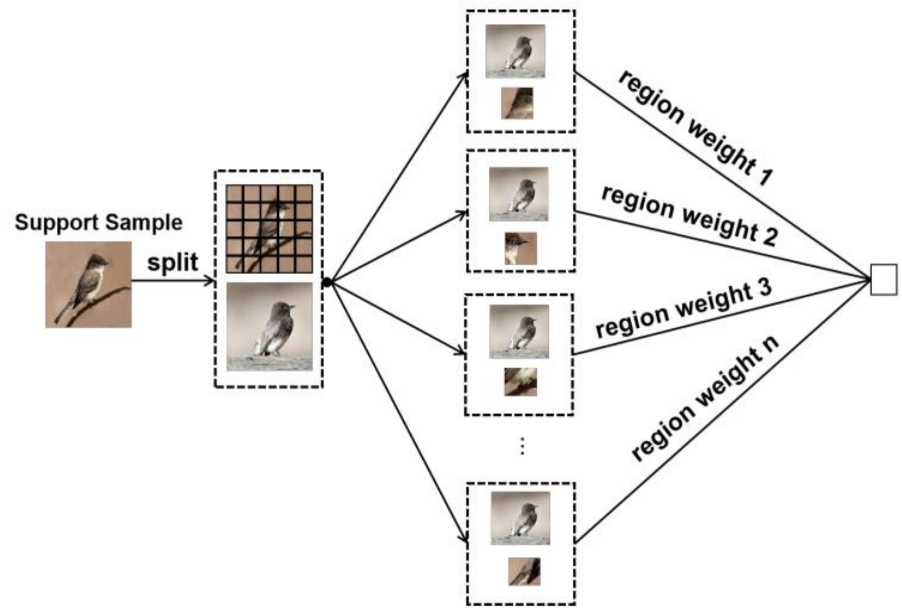
Meta Review from ACM MM 2020: The paper itself does not bring enough insights to the multimedia community. It uses single modality is thus more suitable for vision community.

We plan to submit this paper to AAAI 2020 or TIP, and the codes will be released if the paper is accepted by these conference or journal

Motivation



(a)



(b)

Architecture

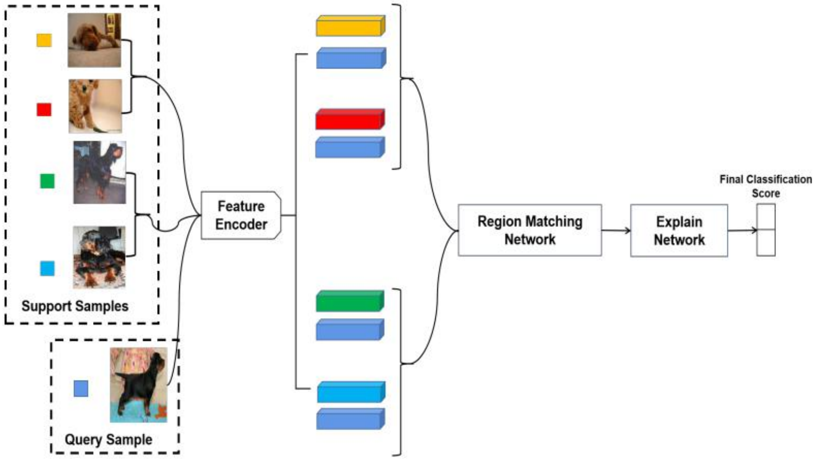


Figure 2: The architecture of 2-way 2-shot

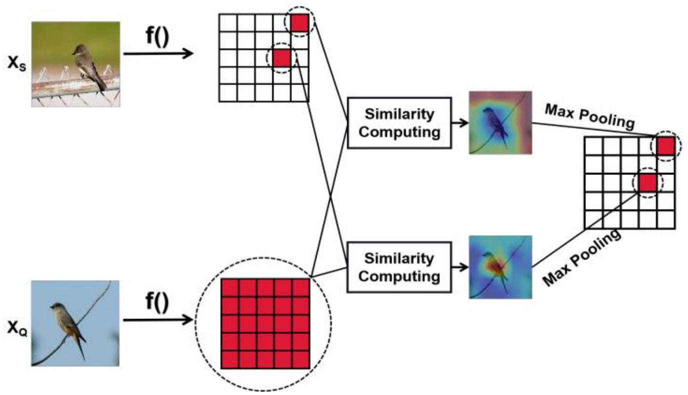


Figure 3: The structure of region matching network for $w = h = 5$, where X_S and X_Q denote support sample and query sample respectively.

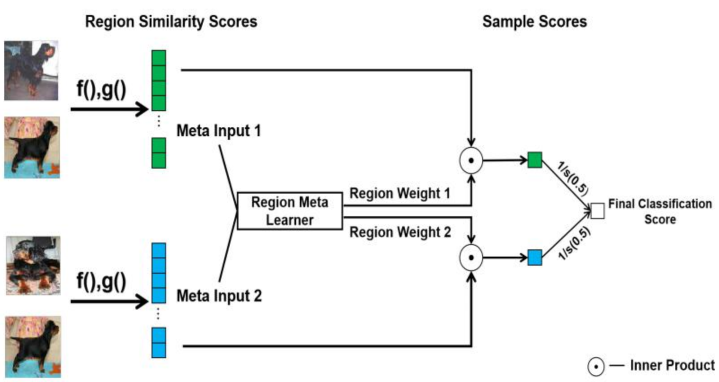


Figure 4: The structure of explain network for 2-shot task(images are from Mini-ImageNet)

Performance

Table 1: Mean accuracies (%) of different methods on the MiniImageNet and CIFAR-FS dataset. Results are obtained over 600 test episodes with 95% confidence intervals. Note that Conv4-n denotes 4-layer convolution network outputting feature maps with n channels. *: [47] uses feature extractor as 6-layer convolution networ with deformable convolution kernel [5]

Model	Backbone	Type	Mini-ImageNet (5-way)		CIFAR-FS (5-way)	
			1-shot	5-shot	1-shot	5-shot
META LSTM [34]	Conv4-32	Meta	43.44±0.77	60.60±0.71	-	-
MAML [7]	Conv4-32	Meta	48.70±1.84	63.11±0.92	58.9±1.9	71.5±1.0
Dynamic-Net [11]	Conv4-64	Meta	56.20±0.86	72.81±0.62	-	-
Dynamic-Net [11]	Res12	Meta	55.45±0.89	70.13±0.68	-	-
SNAIL [30]	Res12	Meta	55.71±0.99	68.88±0.92		
AdaResNet [24]	Res12	Meta	56.88±0.62	71.94±0.57	-	-
MATCHING NETS [43]	Conv4-64	Metric	43.56±0.84	55.31±0.73	-	-
PROTOTYPICAL NETS [40]	Conv4-64	Metric	49.42±0.78	68.20±0.66	55.5±0.7	72.0±0.6
RELATION NETS [41]	Conv4-64	Metric	50.44±0.82	65.32±0.70	55.0±1.0	69.3±0.8
GNN [9]	Conv4-64	Metric	50.33±0.36	66.41±0.63	61.9	75.3
PABN [17]	Conv4-64	Metric	51.87±0.45	65.37±0.68	-	-
TPN [28]	Conv4-64	Metric	52.78±0.27	66.59±0.28	-	-
DN4 [26]	Conv4-64	Metric	51.24±0.74	71.02±0.64	-	-
R2-D2 [2]	Conv4-512	Metric	51.80±0.20	68.4±0.20	65.3±0.2	79.4±0.1
GCR [25]	Conv4-512	Metric	53.21±0.40	72.32±0.32	-	-
PARN [47]	*	Metric	55.22±0.82	71.55±0.66	-	-
RCN	Conv4-64	Metric	53.47±0.84	71.63±0.70	61.61±0.96	77.63±0.75
RCN	Res12	Metric	**57.40±0.86**	**75.19±0.64**	**69.02±0.92**	**82.96±0.67**

Table 2: Mean accuracies (%) of different methods on the CUB-200. Results are obtained over 600 test episodes with 95% confidence intervals. †: Split CUB as [26]. ‡: Split CUB as [4]

Model	Backbone	Type	CUB-200 (5-way)	
			1-shot	5-shot
PCM† [45]	Conv4-64	Metric	42.10±1.96	62.48±1.21
MATCHING NETS† [43]	Conv4-64	Metric	45.30±1.03	59.50±1.01
PROTOTYPICAL NETS† [40]	Conv4-64	Metric	37.36±1.00	45.28±1.03
GNN† [9]	Conv4-64	Metric	51.83±0.98	63.69±0.94
DN4† [26]	Conv4-64	Metric	53.15±0.84	81.90±0.60
RCN†	Conv4-64	Metric	66.48±0.90	82.04±0.58
RCN†	Res12	Metric	**78.64±0.88**	**90.10±0.50**
Baseline++‡ [4]	Res10	Metric	69.55±0.89	85.17±0.50
MAML++(High-End)+SCA‡ [1]	-	Meta	70.46±1.18	85.63±0.66
GPShot(CosSim)‡ [31]	Res10	Meta	70.81±0.52	83.26±0.50
GPShot(BNCosSim)‡ [31]	Res10	Meta	72.27±0.30	85.64±0.29
RCN‡	Conv4-64	Metric	67.06±0.93	82.36±0.61
RCN‡	Res12	Metric	**74.65±0.86**	**88.81±0.57**

Table 4: Mean accuracies (%) of different methods on the Mini-ImageNet and CUB-200(using split criterion as [26]). Results are obtained over 600 test episodes with 95% confidence intervals. Note that the items in region weight of fixed layer are all equal to $\frac{1}{h \times w}$

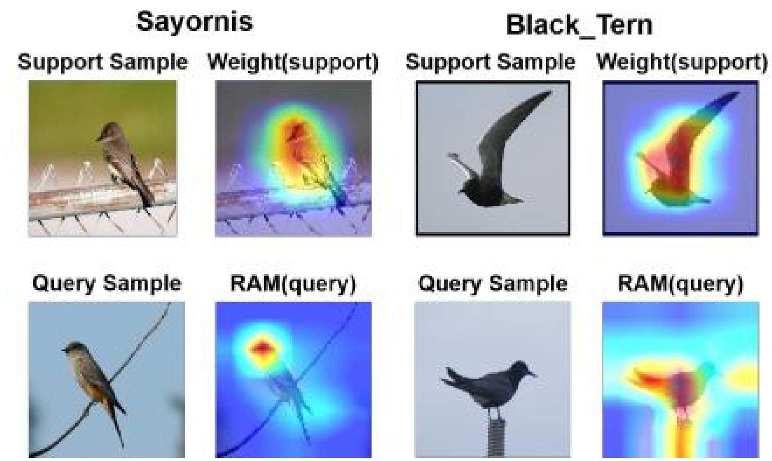
Version	Mini-ImageNet		CUB-200	
	1-shot	5-shot	1-shot	5-shot
Fixed (5×5)	49.30±0.89	55.51±0.71	62.61±1.63	67.26±0.83
Linear (5×5)	55.97±0.86	72.80±0.63	73.23±0.90	88.12±0.56
Meta Learner (5×5)	**57.40±0.86**	**75.19±0.64**	**78.64±0.88**	**90.10±0.50**
Fixed (4×4)	51.79±0.90	57.40±0.70	65.18±1.08	71.65±0.83
Linear (4×4)	55.18±0.84	**73.25±0.64**	75.12±0.89	87.63±0.54
Meta Learner (4×4)	**55.73±0.83**	72.78±0.62	**76.48±0.86**	**87.89±0.57**
Fixed (3×3)	51.51±0.90	56.02±0.70	65.97±1.03	74.59±0.89
Linear (3×3)	**56.50±0.87**	**73.48±0.62**	**76.15±0.87**	**88.10±0.51**
Meta Learner (3×3)	55.41±0.85	72.16±0.68	75.63±0.88	86.96±0.57
Fixed (2×2)	51.58±0.91	57.59±0.70	68.95±1.05	77.64±0.81
Linear (2×2)	**56.03±0.85**	72.23±0.64	73.79±0.85	**87.42±0.57**
Meta Learner (2×2)	55.65±0.83	**72.36±0.64**	**75.79±0.87**	86.64±0.55
Fixed (1×1)	52.22±1.03	57.34±0.75	70.70±0.78	78.43±0.43
Linear (1×1)	54.80±0.86	71.80±0.69	**75.83±0.85**	**86.97±0.53**
Meta Learner (1×1)	**55.40±0.89**	**72.78±0.62**	73.83±0.98	84.77±0.54

Table 3: Mean accuracies (%) of different methods on the Stanford Dogs. Results are obtained over 600 test episodes with 95% confidence intervals.

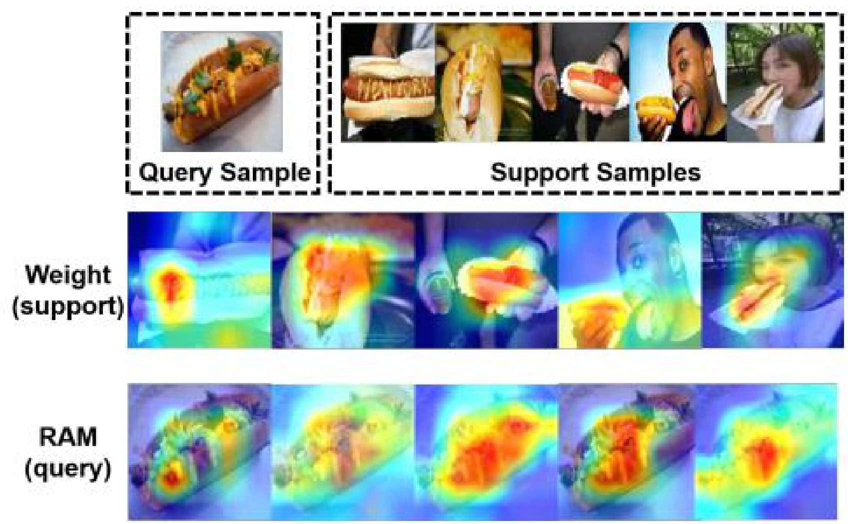
Model	Backbone	Type	CUB-200 (5-way)	
			1-shot	5-shot
PCM [45]	Conv4-64	Metric	28.78±2.33	46.92±2.00
MATCHING NETS [43]	Conv4-64	Metric	45.30±1.03	59.50±1.01
PROTOTYPICAL NETS [40]	Conv4-64	Metric	37.59±1.00	48.19±1.03
GNN [9]	Conv4-64	Metric	46.98±0.98	62.27±0.95
DN4 [26]	Conv4-64	Metric	45.73±0.76	66.33±0.66
RCN	Conv4-64	Metric	54.29±0.96	72.65±0.72
RCN	Res12	Metric	**66.24±0.96**	**81.50±0.58**

VISUALIZATION OF MODEI INTERPRETABILITY

$$RAM = \sum_{i=1}^{h\times w} W_p[i] \cdot k(S^i_{S,Q})$$



(a)



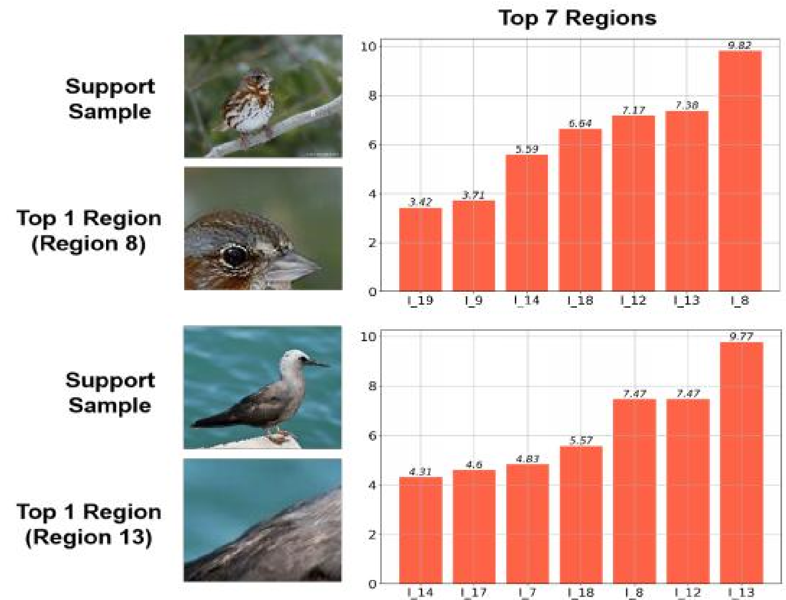
GENERALIZATION AND QUANTIFICATION OF MODEL INTERPRETABILITY

$$f(x, W_{:,j}) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) \tag{6}$$

$$I_j = \int_{\mu_j - 2a}^{\mu_j + 2a} f(x, W_{:,j}) x \mathrm{d}x \qquad a = \frac{1}{M}\sum \sigma_j \tag{7}$$

Algorithm 1 Generalization Method

Input: x_S, $\{x_Q^i\}_{i=1}^{N-1}$
Output: $\{I_j\}_{j=1}^{M}$
$W = []$ is a two-dimensional matrix
$M = 0$
for $x_Q^i \in \{x_Q^i\}_{i=0}^{N-1}$ **do**
  $S_i = m(g(f(x_S), f(x_Q^i)))$
  if $S_i \neq \vec{0}$ **then**
    $W = [W; S_i]$
    $M+ = 1$
  else
    continue
  end if
end for
for $j \in [1, 2, ...M]$ **do**
  $I_j = \int_{\mu_j-2a}^{\mu_j+2a} f(x, W_{:,j}) x \mathrm{d}x$
end for



Explain Class Imbalance Problem by Using Feature Transformation Complexity

Ongoing Project with Prof. Quanshi Zhang in SJTU

I'm not the first author, since the idea is not mine

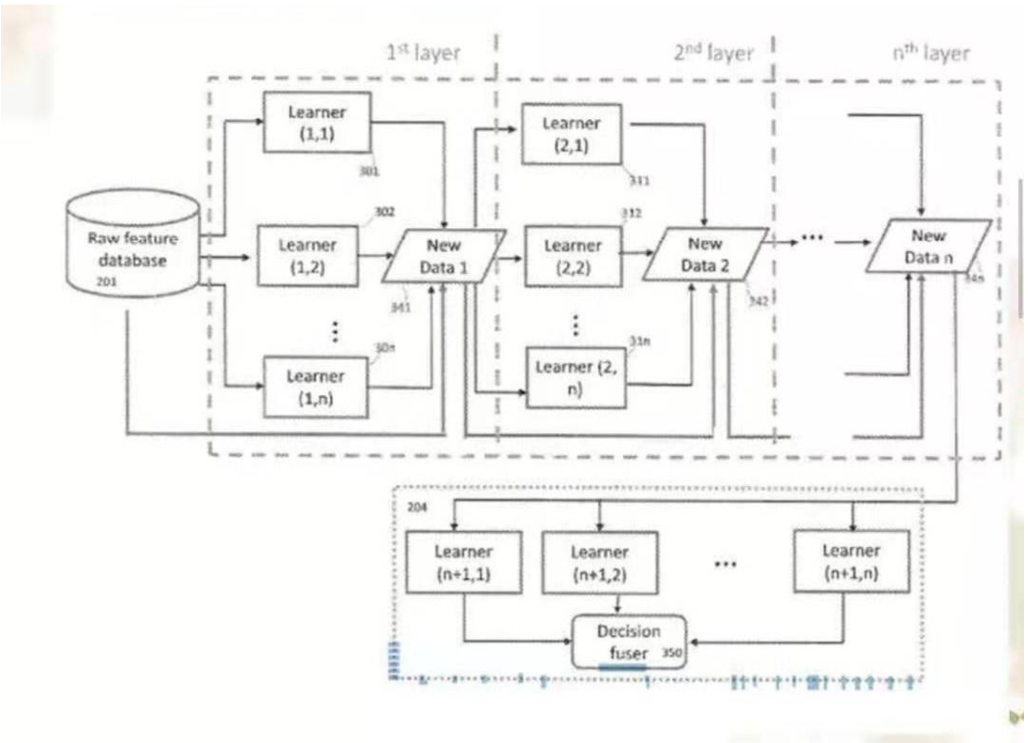
Explain and Improve Few-shot Learning Models by Inversing the Network

Ongoing Project with Prof. Jiebo Luo and Prof. Lixin Duan

I will be the first author

Working Experiences

Data Engineer in Fintell



Patent, IBM 2016, Dr. Changshen Li
Data: ronghui_v7

Reviwer of CVPR VL3 Workshop

I'm on the organizer list of https://www.learning-with-limited-labels.com/organizers

What I Want to Learn

Gradient Compression for Distributed Learning

On Biased Compression for Distributed Learning

Aleksandr Beznosikov* Samuel Horváth Peter Richtárik Mher Safaryan

King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia

February 27, 2020

Abstract

In the last few years, various communication compression techniques have emerged as an indispensable tool helping to alleviate the communication bottleneck in distributed learning. However, despite the fact *biased* compressors often show superior performance in practice when compared to the much more studied and understood *unbiased* compressors, very little is known about them. In this work we study three classes of biased compression operators, two of which are new, and their performance when applied to (stochastic) gradient descent and distributed (stochastic) gradient descent. We show for the first time that biased compressors can lead to linear convergence rates both in the single node and distributed settings. Our *distributed* SGD method enjoys the ergodic rate $\mathcal{O}\left(\frac{\delta L \exp(-K)}{\mu} + \frac{(C+D)}{K\mu}\right)$, where δ is a compression parameter which grows when more compression is applied, L and μ are the smoothness and strong convexity constants, C captures stochastic gradient noise ($C = 0$ if full gradients are computed on each node) and D captures the variance of the gradients at the optimum ($D = 0$ for over-parameterized models). Further, via a theoretical study of several synthetic and empirical distributions of communicated gradients, we shed light on why and by how much biased compressors outperform their unbiased variants. Finally, we propose a new highly performing biased compressor—combination of Top-k

Spare-time Life

Voluntary Teaching, Sri Lanka

Summer School, UC Berkeley

Violist, Ochestra of UESTC

Comments from My Mentors

Overall, Zhiyu is one of the most excellent and diligent students I have ever supervised. I believe his great potentials will continue his driving for excellence in the future, and I know for sure that your prestigious program will boost Zhiyu' s future of success.

-- Prof. Lixin Duan, UESTC, Leader of DIG Lab

During his internship, I found Zhiyu is a warm and friendly student who cooperated well with his teammates. He is always willing to share ideas and organize discussions to find solutions.

-- Dr. Jing Wang, CEO of Fintell Financial Service

Self-summary

Strengths:

1. Self-motivated
2. Quite good at Python
3. Creative (but it sometimes causes blue sky thinking)

Weaknesses & Solutions：

1. Not have a strong mathematics background （Plan to read some papers and books, and I'm highly interested in researching ML problem in the aspect of math）
2. Not good at English （Plan to take GRE test）
3. Time schedule （Force myself to finish the work the day before the deadline）

Thanks for Watching