MTUNet: Few-shot Image Classification with Visual Explanations

Bowen Wang[1,3], Liangzhi Li[1,5], Manisha Verma[1,5], Yuta Nakashima[1,5],
Ryo Kawasaki[2,4], Hajime Nagahara[1,5]
[1]Institute for Datability Science (IDS) [2]Graduate School of Medicine
Osaka University, Japan
[3]bowen.wang@is.ids.osaka-u.ac.jp [4]ryo.kawasaki@ophthal.med.osaka-u.ac.jp
[5]{li, mverma, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

Few-shot learning (FSL) approaches, mostly neural network-based, are assuming that the pre-trained knowledge can be obtained from base (seen) categories and transferred to novel (unseen) categories. However, the black-box nature of neural networks makes it difficult to understand what is actually transferred, which may hamper its application in some risk-sensitive areas. In this paper, we reveal a new way to perform explainable FSL for image classification, using discriminative patterns and pairwise matching. Experimental results prove that the proposed method can achieve satisfactory explainability on two mainstream datasets. Code is available[*].

1. Introduction

Few-shot learning (FSL) is of great significance at least for the following two scenarios [24]: Firstly, FSL can relieve the heavy needs for data gathering and labeling, which can boost ubiquitous use of deep learning techniques, especially for users without enough resources. Secondly, FSL is an important solution for applications in which rare cases matter or image acquisition is costly because of high operation difficulty or ethical issues.

There have been lots of FSL methods [22, 19, 14, 10, 5, 21], most of which are based on the assumption that knowledge can be well extracted from base (seen) classes and transferred to novel (unseen) classes. However, this is not always the case. The knowledge in a pre-trained backbone convolutional neural network (CNN), which computes features of an input image, may sometimes be useless when novel categories have significant visual differences from images of the base categories [26]. What makes matter worse is that we even have no way to see if the visual differences between the base and novel categories are significant *for an FSL model*. This raises one essential question: *Is there any way to see what is actually transferred?*

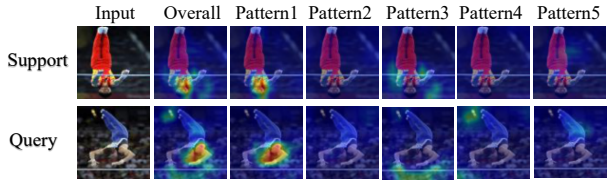

Figure 1. Visualization of each pattern and overall representation for a sampled task in mini-ImageNet.

Actually, in the FSL task, most works [17, 2, 20, 6, 9] only treat the convolutional layer as the image embedding tool, and do not pay attention to the reasons for the extracted features. In this paper, we redesign the mechanism of knowledge transfer for FSL tasks, which offers an answer to the above question. Our approach is inspired by what human beings may do when trying to recognize a rarely seen object. That is, we usually try to find some patterns in the object and match them in a small number of seen examples in our memory.

We adopt a recently-emerged explainable classifier, called SCOUTER [11], and propose a new FSL method, named match-them-up network (MTUNet). MTUNet learns discriminative patterns for a given set of images of the base categories as shown in Figure 1 and uses all these patterns to represent both support and query images. With this representation, pairwise matching scores are computed among the support and query images, based on which the prediction for the query image is done. Both the patterns and the overall representation can be easily visualized to reveal the reason for the matching scores. The main contributions of our work include: 1) a new FSL method that can output visual explanations besides classification results to find potential failures of the method. 2) a new image representation based on filtering original image features, given by a backbone CNN, to keep only informative regions, and relate the visualization with the model decision.

[*]Code is available at https://github.com/wbw520/MTUNet.

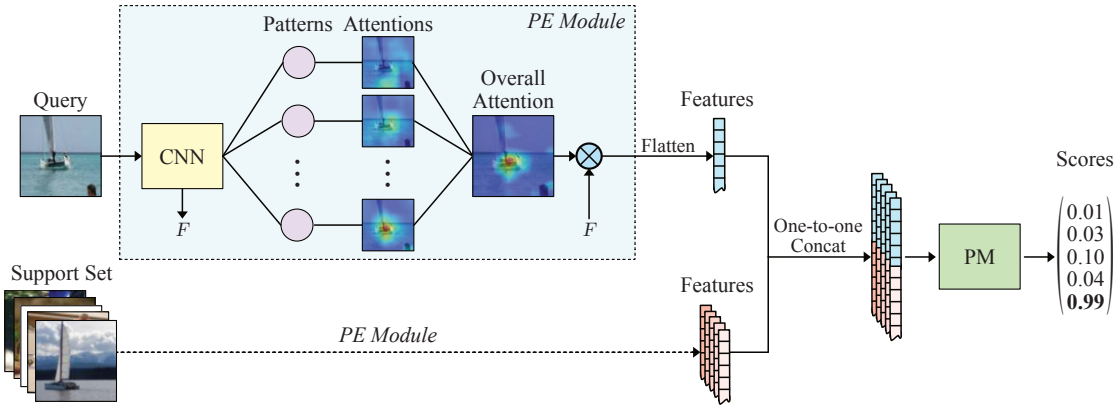

Figure 2. Overall structure of MTUNet. One query is processed by CNN backbone and *pattern extractor* (PE) to provide exclusive patterns and then turned into an overall attention. Query will be concatenated to each support to make a pair for final discrimination through pairwise matching (PM). The dotted line represent each support image undergo the same calculation as query.

2. Methodology

2.1. Problem Definition

This paper addresses an inductive FSL task (*c.f.*, transductive one [3, 8]), in which we are given two disjoint sets $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$ of samples. The former is the base set that includes categories ($\mathcal{C}_{\text{base}}$) with many labeled images. The latter is the novel set and include categories ($\mathcal{C}_{\text{novel}}$) with a few labeled images. $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ are disjoint. The FSL task is to find a mapping from a novel image x into the corresponding category y.

The literature typically uses the K-way N-shot episodic paradigm for training/evaluating FSL models. For each episode, we sample two subsets of $\mathcal{D}_{\text{base}}$ for training, namely, *support set* $\mathcal{S} = \{(x_i, y_i)|i = 1, \ldots, K \times N\}$ and *query set* $\mathcal{Q} = \{(x_i^{\text{q}}, y_i^{\text{q}})|i = 1, \ldots, K \times M\}$. These images are of the same K categories in $\mathcal{C}_{\text{base}}$, and we sampled the same numbers of images (N images for the support set and M images for the query set). An FSL model is trained so that it can find a match between images in $\mathcal{Q}$ (with abuse of notation) and $\mathcal{S}$. The image in $\mathcal{Q}$ is classified as the category of the matched image in $\mathcal{S}$.

2.2. Overview

The overall process is illustrated in Figure 2. In each episode, we extract feature map $F = f_\theta(x) \in \mathbb{R}^{c\times h\times w}$ from image x in both $\mathcal{S}$ and $\mathcal{Q}$ using backbone convolutional neural network f_θ, where θ is the set of learnable parameters. F is then fed into the *pattern extractor* (PE) module, f_ϕ, with learnable parameter set ϕ. This module gives attention $A = f_\phi(F) \in \mathbb{R}^{z\times l}$ over F. Our *pair matching* (PM) module uses an MLP to compute the score of query image $x^q \in \mathcal{Q}$ belonging to the category of x's in $\mathcal{S}$.

PE plays a major role in the FSL task. PE is designed to learn a transferable attention mechanism. This ends up in finding common patterns that are shared among different episodes sampled from $\mathcal{D}_{\text{base}}$. Consequently the patterns are shared also among $\mathcal{D}_{\text{novel}}$ given that $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$ are from similar domains.

2.3. Pattern Extractor

The basic idea of PE is to find common patterns through the self-attention mechanism. Input feature map F is firstly fed into a 1×1 convolution layer followed ReLU nonlinearity to squeeze the dimensionality of F from c to d. Spatial dimensions of the squeezed features are flattened to form $F' \in \mathbb{R}^{d\times l}$, where $l = hw$. To maintain the spatial information, position embedding P [25, 13, 11] is added to the features, *i.e.*, $\tilde{F} = F' + P$.

The self-attention mechanism gives the attention over F for the spatial dimension by the dot-product similarity between a set of z learned patterns $W \in \mathbb{R}^{z\times d}$ (z is the number of the patterns) and $\tilde{F}$ after nonlinear transformations g_{Q} and g_{K}. PE repeats this process with updating the pattern with a gated recurrent unit (GRU) to refine the attention. That is, given

$$g_{\text{Q}}(W^{(t)}) \in \mathbb{R}^{z\times d}, \quad g_{\text{K}}(\tilde{F}) \in \mathbb{R}^{d\times l}, \tag{1}$$

for the t-th repetition, the attention is given using certain normalization function ξ by

$$\bar{A}^{(t)} = g_{\text{Q}}(W^{(t)}) g_{\text{K}}(\tilde{F}) \quad \in (0, 1)^{z\times l} \tag{2}$$

$$A^{(t)} = \xi(\bar{A}^{(t)}). \tag{3}$$

Patterns $W^{(t)}$ is updated T times (*i.e.*, $t = 1, \ldots, T$) by

$$U^{(t)} = A^{(t)} {F'}^{\top} \tag{4}$$

$$W^{(t+1)} = \text{GRU}\,(U^{(t)}, W^{(t)}). \tag{5}$$

PE adopts a different normalization strategy from SCOUTER. Let $\text{Softmax}_{\text{R}}(X)$ and $\sigma(X)$ be softmax over

respective row vectors of matrix X and sigmoid respectively. SCOUTER normalizes the attention map only over the flattened spatial dimensions, *i.e.*,

$$A^{(t)} = \sigma(\bar{A}^{(t)}). \tag{6}$$

This allows finding multiple patterns in a single image. MTUNet further modulates this map by

$$A^{(t)} = \sigma(\bar{A}^{(t)}) \odot \text{Softmax}_{\text{R}}(\bar{A}^{(t)}), \tag{7}$$

which suppresses weak attention over different patterns at the same spatial location, where $\odot$ is the Hadamard product. This enforces the network to find more specific yet discriminative patterns with fewer correlations among them and thus ends up with more pinpoint attentions. The learned patterns can be more responsive in different images with this modulation as an attention map only responds to a single pattern that does not include its peripheral region.

The input image is finally described by the overall attention corresponding to the extracted patterns, given by

$$V = \frac{1}{z} A^{(T)} F \mathbf{1}_z, \tag{8}$$

where $\mathbf{1}_z$ is the row vector with all z elements being 1. $A^{(T)}$ is reshaped from l into the same spatial structure as F. V will then undergo an average pooling among spatial dimension and only keep the channel dimension c.

2.4. Pairwise Matching

An FSL classification can be solved by finding the membership of the query to one of the given support images. Learnable distances is a popular choice for the metric learning-based FSL methods [10, 5, 21]. We use a learnable distance with an MLP.

Let V^{q} and V_{kn} be features of query image $x^{\text{q}} \in \mathcal{Q}$ and support image $x_{kn} \in \mathcal{S}$ respectively, where the subscripts k and n stand for the n-th image of category k. For $n > 1$, the average over the n images are taken to generate representative feature $\bar{V}_k$; otherwise (*i.e.*, $n = 1$), $\bar{V}_k = V_{k1}$. For computing the membership score s of query image x^{q} to category k, we use MLP f_γ with learnable parameters γ:

$$s(x^{\text{q}}, \mathcal{S}_k) = \sigma(f_\gamma([V^{\text{q}}, \bar{V}_k])), \tag{9}$$

where $[\cdot, \cdot]$ is concatenation of two vectors for the one-to-one pair and $\mathcal{S}_k \subset \mathcal{S}$ contains images of category k. x^{q} is classified into the category with maximum s over $\mathcal{S}_k$ for $k = 1, 2, \ldots, K$.

For $\mathcal{Q}$ and $\mathcal{S}$ sampled from $\mathcal{D}_{\text{base}}$ for each episode, we train the model with the binary cross-entropy loss:

$$L = -\sum_{(x^{\text{q}}, y^{\text{q}}) \in \mathcal{Q}} {y^{\text{q}}}^\top \log(\bar{s}(x^{\text{q}}, \mathcal{S})), \tag{10}$$

where $\bar{s}(x^{\text{q}}, \mathcal{S}) = (s(x^{\text{q}}, \mathcal{S}_1), \ldots, s(x^{\text{q}}, \mathcal{S}_K))^\top$.

Table 1. Average accuracy of 10000 sampling 5-ways task on test set of mini-ImageNet and tired-ImageNet.

	mini-ImageNet		tiered-ImageNet	
Approach	**One shot**	**Five shots**	**One shot**	**Five shots**
MetaLSTM [16]	43.44±0.77	60.60±0.71	-	-
MAML [4]	48.70±1.84	63.11±0.92	51.67±1.81	70.30±0.08
ProtoNet [19]	49.42±0.78	68.20±0.66	53.31±0.20	72.69±0.74
Meta SGD [12]	50.47±1.87	64.03±0.94	62.95±0.03	79.34±0.06
Reptile [15]	49.97±0.32	65.99±0.58	48.97±0.21	66.47±0.21
R2-D2 [1]	51.20±0.60	68.20±0.60	-	-
RelationNet [21]	52.48±0.86	69.83±0.68	54.48±0.93	71.32±0.78
SimpleShot(UN) [23]	57.81±0.21	**80.43±0.15**	64.35±0.23	**85.69±0.15**
LEO [18]	**61.76±0.08**	77.59±0.12	**66.33±0.05**	81.44±0.09
MTUNet+ResNet-18	55.03±0.49	70.22±0.35	61.27±0.50	77.82±0.41
MTUNet+WRN	56.12±0.43	71.93±0.40	62.42±0.51	80.05±0.46

3. Experiments

3.1. Experimental Setup

We evaluate our approach on two commonly-used datasets, mini-ImageNet [22] and tiered-ImageNet [17]. They are split into train, validation and test sets. We evaluate MTUNet on 10,000 episodes of 5-way classification created by first randomly sampling 5 categories from $\mathcal{D}_{\text{base}}$ and then sampling support and query images of these categories with $N = 1$ or 5 and $M = 15$ per category. We report the average accuracy over $K \times M = 75$ queries in the 10,000 episodes and the 95% confidence interval. We employ two CNN architectures as our backbone f_θ, which are namely WRN-28-10 [27] and ResNet-18 [7].

For pre-training of PE module, we used the same parameter setting as [11]. The number z of the patterns is empirically set to $1/10$ of the train set categories, which is 7 in mini-ImageNet. The importance of this choice is discussed in Section 3.4. For training the whole MTUNet, the learnable parameters in backbone CNNs and PE are frozen. In a single epoch of training, we sample 1,000 episodes of 5-way tasks. The model is trained for 20 epochs with an initial learning rate 10^{-3}, which is divided by 10 at the 10-th epoch. We use the model with the best performance with 2,000 episodes sampled from the validation set.

3.2. Results

For comparison, we exclude ones in semi-supervised and transductive paradigms [8, 3], which use the statistics of the novel set. We report our best model by randomly sampling 10,000 1-shot and 5-shot tasks over the test set in Tables 1.

3.3. Explainability

MTUNet is designed to be explainable in two different aspects. Firstly, MTUNet's decision is based on certain combinations of learned patterns. These patterns are localized in both query and support images through $A^{(T)}$, which can be easily visualized. Secondly, thanks to the one-to-one matching strategy formulated as a binary classification

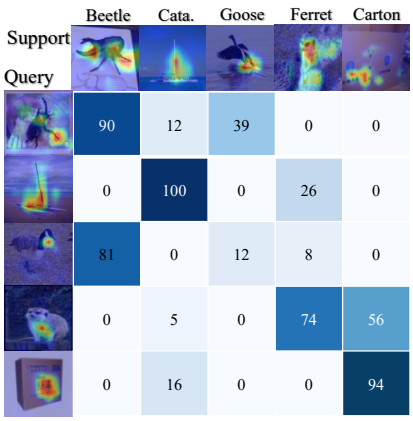

Figure 3. Matching point matrix of one sampled task in mini-ImageNet. Row and column are consisted with the overall attention visualization for support and query of each category.

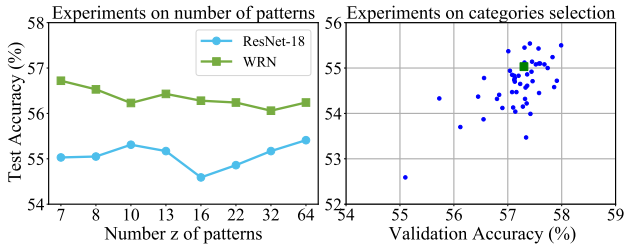

Figure 4. Experiments on the numbers of patterns and the categories selection.

problem in Eq. (9), the distributions (or appearances) of learned patterns in query and support images give a strong clue on MTUNet's matching score s.

Pattern-based visual explanation. Figure 1 shows a pair of support and query images in a 5-way task in mini-ImageNet. The pairs are of category `horizontal bar`. The second column shows the visualization of overall attention, given by

$$A' = \frac{1}{z} A^{(T)} \mathbf{1}_z. \tag{11}$$

For support image, we can see that the visualization of pattern 1 identifies the part of the human body (head), and pattern 3 appears around the hands grabbing the horizontal bar. For the query image, patterns 1 and 3 respond in a similar way to the support image. Patterns 4 and 5 appear in the background and around other parts of the body, however their responses are relatively weak. Patterns 1 and 3 may responsible for human heads and hands grabbing the horizontal bar, lead to successful classification.

Visualization of pairwise matching scores Figure 3 shows the visualization of the pairwise matching score. The output for each pair is a value between 0 to 1 due to the sigmoid function, whereas the scores are shown in percentage in the figure. The combination of the support and query images of `catamaran` makes the full score (100%). The visualization of overall attention covers the hulls, especially the masts, in both images, which are the main characteristics of this category. Surprisingly, the query image for `goose` gets 81% for the support image for `beetle`. This may suggest that one of the patterns responds to black regions and this pattern is solely used as the clue of `goose`. This is a negative result for the FSL tasks because it means the model does not generalize, but clearly demonstrates MTUNet's explainability on the relationship between visual patterns and the pairwise matching scores.

3.4. Discussion

The number z of patterns. The number of patterns can be a crucial factor for MTUNet. The test accuracies are shown in Figure 4 for 5-way 1-shot tasks in 10,000 sampled episodes over $\mathcal{D}_{\text{test}}$. The horizontal axis represents the number of patterns and the vertical axis represents the average accuracy. Interestingly, the results show no clear tendency with respect to z. In general, tuning over z may help gain performance, but its impact is not significant.

Selection of categories for training PE. Our PE module is supposed to learn common visual patterns. We use images of a certain subset of categories in $\mathcal{C}_{\text{base}}$ to learn such patterns in our experiments. To clarify the impact of the choice of the subset, we randomly sample seven categories in $\mathcal{C}_{\text{base}}$ of mini-ImageNet for 50 times and use the corresponding images for training PE on top of ResNet-18. Figure 4 shows the scatter plot of the validation accuracies and corresponding test accuracies, which has Pearson's correlation coefficient of 0.64. This leads to the conclusion that, at least for mini-ImageNet, we can use the validation set to find the better choice. The green square in the plot is the choice that we adopted in our experiments.

4. Conclusion

In this paper, we propose MTUNet designed for explainable FSL. We achieved comparable performance on two benchmark datasets and qualitatively demonstrated its strong explainability through patterns in images. The approach adopted in our model might be analogous to human beings as they usually try to find shared patterns when making a match between images of an object that one has never seen before. This can be advantageous as the explanation given by MTUNet can provide an intuitive interpretation of what the model actually learns.

Acknowledgements This work was supported by Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), "Innovative AI Hospital System" (Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN)). This work was also supported by JSPS KAKENHI Grant Number 19K10662 and 20K23343.

References

[1] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *ICLR*, 2019. 3

[2] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *IEEE CVPR*, pages 8680–8689, 2019. 1

[3] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *ICLR*, 2020. 2, 3

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017. 3

[5] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *ICLR*, 2018. 1, 3

[6] Yuxia Geng, Jiaoyan Chen, Zhiquan Ye, Wei Zhang, and Huajun Chen. Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. *SWJ*, 2020. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 3

[8] Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. *arXiv preprint arXiv:2006.03806*, 2020. 2, 3

[9] Leonid Karlinsky, Joseph Shtok, Amit Alfassy, Moshe Lichtenstein, Sivan Harary, Eli Schwartz, Sivan Doveh, Prasanna Sattigeri, Rogerio Feris, Alexander Bronstein, et al. StarNet: towards weakly supervised few-shot detection and explainable few-shot classification. *arXiv preprint arXiv:2003.06798*, 2020. 1

[10] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *IEEE CVPR*, pages 11–20, 2019. 1, 3

[11] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. SCOUTER: Slot attention-based classifier for explainable image recognition. *arXiv preprint arXiv:2009.06138*, 2020. 1, 2, 3

[12] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to learn quickly for few-shot learning. *ICML*, 2017. 3

[13] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 2020. 2

[14] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *ICLR*, 2018. 1

[15] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 3

[16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *ICLR*, 2017. 3

[17] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *ICLR*, 2018. 1, 3

[18] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *ICLR*, 2019. 3

[19] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017. 1, 3

[20] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explain and improve: Cross-domain few-shot-learning using explanations. *arXiv:2007.08790*, 2020. 1

[21] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE CVPR*, pages 1199–1208, 2018. 1, 3

[22] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016. 1, 3

[23] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 3

[24] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020. 1

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE CVPR*, pages 1492–1500, 2017. 2

[26] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*, 2020. 1

[27] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3