MA^3: Model Agnostic Adversarial Augmentation for Few Shot learning

Rohit Jena rjena@cs.cmu.edu
Shirsendu Sukanta Halder shirsenh@cs.cmu.edu
Katia Sycara katia@cs.cmu.edu

Carnegie Mellon University

Abstract

Despite the recent developments in vision-related problems using deep neural networks, there still remains a wide scope in the improvement of generalizing these models to unseen examples. In this paper, we explore the domain of few-shot learning with a novel augmentation technique. In contrast to other generative augmentation techniques, where the distribution over input images are learnt, we propose to learn the probability distribution over the image transformation parameters which are easier and quicker to learn. Our technique is fully differentiable which enables its extension to versatile data-sets and base models. We evaluate our proposed method on multiple base-networks and 2 data-sets to establish the robustness and efficiency of this method. We obtain an improvement of nearly 4% by adding our augmentation module without making any change in network architectures. We also make the code [1] readily available for usage by the community.

1. Introduction

Supervised learning algorithms have demonstrated tremendous success in a multitude of tasks both high-level like classification [21], detection [18], etc and also in low-level tasks such as segmentation [14] after the explosion of deep neural networks. However, the same statement cannot be made for situations where the model is expected to generalize in the absence of densely available labels. This is unlike humans, who generalise in an incremental manner to novel classes by observing only a few number of examples [12]. The importance of a learning model that improves on unseen examples on gathering more experience is instrumental in almost all practical problems where annotating labels is either not scalable or unavailable due to safety or privacy issues.

Motivated by the aforementioned issues, recent approaches to generalize learning models range from weakly-supervised learning [3], transfer learning [25], domain adaptation techniques [15], data augmentation [20], incremental learning [17] and task based few shot learning [26, 23, 22, 24]. Few-shot classification aims to accommodate to novel classes unseen during training by just using a few examples during test time. This is unlike fine-tuning, where the classifier uses a previously learnt representation and tunes its parameters to maximize accuracy over the new data. The problem with fine tuning is that the classifier would most likely overfit to the new data when it is given as few as five examples. In this work, we take inspiration from humans in the sense that in order for registration, we infer the scene from different perspectives and then are able to generalize in similar future settings. We present a novel method for end-to-end differentiable data augmentation technique inspired by Spatial Transformer Networks [8] and inference technique for single and few-shot learning scenarios. Our contributions are as follows:

1. We propose a theory for a new data-augmentation technique inspired from projective transformations in the 3D camera pinhole model.
2. We demonstrate an algorithm that estimates the data augmentation parameters in an end-to-end neural network model to generalize under a multi-class *k*-shot classification framework.
3. We present analysis of our proposed algorithm using 3 recent few-shot learning paradigms and establish the efficiency of our method for one-shot and few-shot learning on two versatile datasets.

The rest of the paper is as follows. Section 2 presents some of the previous works in literature pertaining to learning with limited labels and data augmentation techniques. Section 3 describes our method in detail followed by Section 4 which shows detailed analysis and comparison. The final Section 5 contains concluding remarks and discussions about scope for future work.

2. Related Work

Few-shot learning: Lake *et al*. [10] propose a generative model and infer handwritten characters from latent strokes in new characters. Ravi and Larochelle [16] use a LSTM-based *meta-learner* that captures short-term knowledge particular to a task and long-term knowledge common

[1] https://github.com/rohitrango/STNAdversarial

to all tasks. ProtoNets [22] learn a representation based metric space and perform classification using the "prototypes" (class means) of each class. Vinyals *et al*. [26] propose a network called Matching Networks that learns the mapping between a small labelled support set and an unlabelled example. The principle that the testing and training conditions should match is used for the training procedure. Few-shot learning has also been explored in the context of meta-learning by Finn *et al*. [6] where they propose an algorithm for fast adaptation of networks on versatile tasks and demonstrate their effectiveness on one-shot learning tasks. Finn *et al*. [7] further explore the task of one-shot learning for a robot under the framework of meta-learning combined with imitation learning from visual demonstrations. Meta learning and transfer learning was combined by [23] to propose an efficient learning curriculum which they name *hard-task meta batch scheme* that improves the convergence and accuracy.

Data augmentation: Antoniou *et al*. [2] were the first to demonstrate improved performances on meta-learning tasks using data augmentation techniques. They do so by generalizing the model to generate class-agnostic data samples. Zhang *et al*. [27] approach the problem of few-shot learning using a unified adversarial generator that is capable of learning sharper boundaries for supervised few-shot and semi-supervised few-shot scenarios as well. This is facilitated by making the GAN generate fake data that provides additional examples for training. Our method is also based on adversarial training but instead of directly generating augmented examples for training, we generate the parameters for transforming the input to learn a robust classifier. The closest work compared to ours is [5] where they use a search algorithm to search the best policy for augmenting a single sample in a mini-batch. The policies consist of sub-policies consisting of either rotation, translation or shearing functions. However, the method is not tested in few-shot settings and the use of reinforcement learning can be unstable with an evolving reward function. Our work is different in the sense that instead of considering these image processing functions independently, we use an adversarial scheme to learn the complete affine transform matrix elements which provides us with better generalization. We also show that a variant which predicts the parameters independently doesn't perform as well as our method.

3. Method

Our model takes inspiration from how humans observe novel objects - they don't just register one "snapshot" of the object, but rather take a look from multiple coherent perspectives. Although this may not be possible given that we do not have images of the same object taken from different perspectives, we can approximate it by assuming that the object is placed far away from the camera (i.e. $z \approx z_0 \gg 1$).

Consider a 3D point of an object in homogeneous coordinates $\begin{pmatrix} x & y & z & 1 \end{pmatrix}^T$ and its 2D projection into the image plane $\begin{pmatrix} u & v & 1 \end{pmatrix}^T$. Without loss of generality, assume that $R = I, t = 0$ to get $u_1 = x/z_0, v_1 = y/z_0$.
Consider a slight change of roll (γ), yaw (α) and pitch (β) where $\|\gamma\|, \|\alpha\|, \|\beta\| \ll 1$, and a small change in translation $\boldsymbol{t}$ such that $\|\boldsymbol{t}\| \ll 1$. Plugging these formulae into the rotation matrix and using Taylor expansion (ignoring third order terms and higher), we have:

$$R = \begin{bmatrix} 1 - \frac{\alpha^2}{2} - \frac{\beta^2}{2} & \beta\gamma - \alpha & \beta + \alpha\gamma \\ \alpha & 1 - \frac{\alpha^2}{2} - \frac{\gamma^2}{2} & \alpha\beta - \gamma \\ -\beta & \gamma & 1 - \frac{\beta^2}{2} - \frac{\gamma^2}{2} \end{bmatrix}$$

and

$$t = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$$

The new point in the image plane corresponding to the original 3D coordinate is:

$$u_2 = \frac{(1 - \frac{\alpha^2}{2} - \frac{\beta^2}{2})x + (\beta\gamma - \alpha)y + (\beta + \alpha\gamma)z_0 + t_x}{-\beta x + \gamma y + (1 - \frac{\beta^2}{2} - \frac{\gamma^2}{2})z_0 + t_z}$$

Since we assume it to be a distant object, and the values of α, β, γ are relatively small, the denominator can be simplified using binomial expansion

$$\frac{1}{z_0(1-\delta)} \approx \frac{1+\delta}{z_0}$$

where $\delta = \frac{\beta^2}{2} + \frac{\gamma^2}{2} + \frac{\beta x}{z_0} - \frac{\gamma y}{z_0} + \frac{t_z}{z_0}$ The new point on the image plane is approximated as

$$u_2 \approx (1+\delta)\left[\left(1 - \frac{\alpha^2}{2} - \frac{\beta^2}{2}\right)\frac{x}{z_0} + (\beta\gamma - \alpha)\frac{y}{z_0} + \left(\beta + \alpha\gamma + \frac{t_x}{z_0}\right)\right]$$

and

$$v_2 \approx (1+\delta)\left[\alpha\frac{x}{z_0} + \left(1 - \frac{\gamma^2}{2} - \frac{\beta^2}{2}\right)\frac{y}{z_0} + \left(\alpha\beta - \gamma + \frac{t_y}{z_0}\right)\right]$$

Substituting the values of u_1, v_1 we get

$$\begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1+\delta_1 & \delta_2 & \delta_3 \\ \delta_4 & 1+\delta_5 & \delta_6 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}$$

where $\|\delta_i\| \ll 1, \forall i \in \{1..6\}$ We approximate the distortion in rotation and translation using an affine transform of the given form, which encourages only slight deviation from the identity transform. The values of the parameters δ_i can be determined using an adversary that detects the distortions that the model hasn't generalized to. This is the core idea

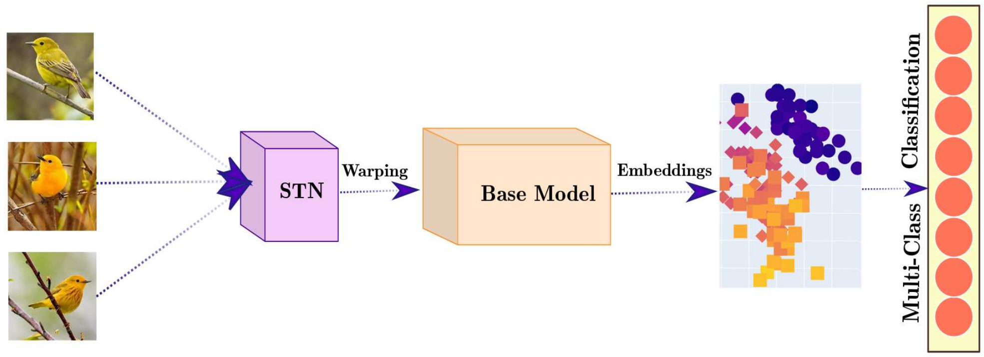

Figure 1. Proposed augmentation module to few shot learning

which forms the basis of generalization to unseen examples. We use Spatial Transformer Networks (STN) which are end-to-end differentiable spatial manipulators. STN computes parameters of the spatial manipulation rather than the manipulated image itself, making it easier to learn a few parameters and perform powerful spatial transformations. They are generally used as a starting module to output a canonical version of an image that can be used as input to a classifier. However, we use it in an adversarial manner by backpropagating through the Cross Entropy loss of the few-shot learner. Learning the trend of the parameters δ_i is simpler and quicker than GANs that learn the data distribution over entire images in response to a noise signal or other support images. We show that this form of augmentation to an image is more effective than applying standard augmentations like random rotations, translations and scaling. At every epoch, the few shot learner processes a batch of support and query examples. The few shot network minimizes the classification loss on the query examples given the support examples. The Transformer takes gradients with respect to the support images to maximize the classification loss on the query images. Let the transformer be a function f parameterized by ϕ and the few shot learner is a function g parameterized by θ. Let $S = \{s_1, s_2, \dots s_n\}$ be the support dataset and $Q = \{q_1, q_2 \dots q_m\}$ be the query dataset. The optimization problem becomes:

$$\max_{\phi} \min_{\theta} \sum_{i=1}^{m} L(g_\theta(q_i | f_\phi(s_1), f_\phi(s_2) \dots f_\phi(s_n)))$$

To make sure that the Transformer doesn't deviate from the identity transform, we apply a regularization term that penalizes deviation from the identity affine transform. The regularization is given by the following term:

$$L_{reg}(f_\phi(s)) = \left\| \begin{bmatrix} a_1(s) & a_2(s) & a_3(s) \\ a_4(s) & a_5(s) & a_6(s) \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right\|^2$$

The modified optimization problem becomes:

$$\max_{\phi} \min_{\theta} \sum_{i=1}^{m} L(g_\theta(q_i | f_\phi(s_1), f_\phi(s_2) \dots f_\phi(s_n))) - \lambda \sum_{j=1}^{n} L_{reg}(f_\phi(s_j))$$

where λ is a hyperparameter. Note that regularization plays an important role, because without any regularization the STN can morph the images to have unrecognizable features and hence maximizing the classification loss and not allowing the classifier to learn useful features. Without explicit regularization, the parameters of the affine matrix predicted by the STN will also violate the assumption about the magnitudes of the δ parameters. This does occur in our experiments when we set $\lambda = 0$, the accuracy over the validation set decreases because the classifier failed to learn good features during training.

4. Experiments

To analyse the effect of adversarial Spatial Transformer Networks, we test our training framework on the Omniglot [11] and MiniImageNet [19] datasets. We show that our method is base-model agnostic by testing on 3 different methods - Prototypical Networks [22], Matching Networks [26] and Model-Agnostic Meta Learning (MAML) [6] frameworks for few shot learning. We observe that all baselines have very high accuracy on the Omniglot dataset, and adding an STN improves the results only marginally. Therefore, we show results for Omniglot only with Prototypical Networks. However, the improvements in accuracy for MiniImageNet are significant and we test our module with all the three baselines.

Prototypical networks received some concerns about reproducibility in results [4], [13], [1]. To provide consistent results for all methods, we use the code provided by [9] and incorporate our module into the code.

In standard classification tasks, the training data is augmented and the validation data is not augmented. We follow the same procedure, we augment the meta-train (or support) examples and do not augment the meta-validation (or query) examples while training. During test time, the STN is disabled for both support and query examples. To avoid potential data distribution shift between the support examples encountered during the training phase and validation phase, we apply a dropout on the output of the STN to retain some of the support images (by randomly selecting images and setting their affine matrix to identity). The dropout value is fixed to 0.5 and the values of λ are obtained using a coarse grid search on a log-scale and a finer grid search on a linear scale after choosing the best interval from the coarse search. The first baseline does not use any data augmentation. The second baseline uses standard data augmentation like random scaling, translation, and rotation. However, unlike random data augmentation, our method outputs parameters by an adversarial STN. The STN outputs the values of rotation θ, translation p_x, p_y and scale s and the affine matrix is constructed as:

$$A = \begin{bmatrix} s\cos(\theta) & -s\sin(\theta) & p_x \\ s\sin(\theta) & s\cos(\theta) & p_y \end{bmatrix}$$

The values are bounded to $\theta \in [-\theta_0, \theta_0]$, $s \in [1-\epsilon_s, 1+\epsilon_s]$ and $p_x, p_y \in [-T, T]$ using tanh activations and appropriate scaling. For all experiments, we set $\theta_0 = \pi$, $\epsilon_s = 0.1$, and $T = 0.1\max(H, W)$, where H, W are the height and width of the images.

Table 1. Quantitative comparison of our method with baseline methods on Omnigot [11] dataset. The base network used in this scenario is ProtoNets [22] with $h_{dim} = 128$ and $\gamma = 0.5$. The comparisons provided in both the tables are with vanilla-baseline method, baseline method with commonly used augmentation techniques, our proposed method with no constraint/regularization on the transformation parameters and our method with constrained parameters.

Classification Task	*Baseline*	*Baseline* (with standard aug.)	*Ours* ($\lambda = 0$)	*Ours*
20 way, 5 shot	98.70%	**98.89%**	94.25%	98.80%
20 way, 1 shot	95.9%	**96.09%**	80.70%	95.97%
5 way, 5 shot	99.62%	99.62%	99.40%	**99.67%**
5 way, 1 shot	98.42%	98.60%	96.40%	**98.61%**

The improvements on Prototypical Networks for Omniglot dataset (Table 1) are not very significant because the baselines already learn features which are general enough to perform well on this easy dataset. However, miniImageNet is a dataset with more variance and would require a classifier to learn complex features to perform well. Our method bumps the performance of the base classifiers by as much as 3.8% without requiring any change to the model architecture, thereby learning better features than that are learnt without the adversarial augmentation (Table 2). Expectedly, our method fails to generalize in the absence of regularization as the STN exploits the freedom of choosing the affine matrix by performing transformations which produce images that are very far from the original data distribution and are often degenerate (for example, excessively zoomed images can result in the image being just a single color). These images hinder the actual learning of the classifier and the accuracy drops significantly below the baseline method. This clearly reinforces our hypothesis regarding the importance of regularization while estimating the transformation parameters. Baseline with standard data augmentation performs better than the baseline in most cases, but the improvement is not consistent (see table 2 - 5 way, 5 shot in ProtoNets and 5 way, 1 shot in MAML).

Table 2. Quantitative Comparison of our method with baseline methods on miniImageNet [19] dataset. The first table contains results using ProtoNets, followed by MAML [6] followed by Matching Nets [26] as the base network.

Classification Task (ProtoNets [22])	*Baseline*	*Baseline* (with standard aug.)	*Ours* ($\lambda = 0$)	*Ours*
5 way, 5 shot	66.6%	70.2%	58.8%	**70.4%**
5 way, 1 shot	51.4%	49.8%	36.2%	**52.8%**

Classification Task (MAML [6])	*Baseline*	*Baseline* (with standard aug.)	*Ours* ($\lambda = 0$)	*Ours*
5 way, 5 shot	65.9%	66.3%	57.9%	**67.0%**
5 way, 1 shot	47.3%	47.3%	32.1%	**48.2%**

Classification Task (Matching Nets [26])	*Baseline*	*Baseline* (with standard aug.)	*Ours* ($\lambda = 0$)	*Ours*
5 way, 5 shot	59.8%	61.4%	47.8%	**62.0%**
5 way, 1 shot	47.0%	48.4%	34.2%	**50.8%**

5. Conclusion

In this paper, we introduced MA^3, a model-agnostic adversarial augmentation technique for few shot learning. The method is inspired by an approximate model of how humans "cheat" by observing a novel object from various perspectives. We show that the model can be approximated using an affine transform, and Spatial Transformer Networks naturally fit into the equation by predicting affine transforms that the classifier is not robust to. Experiments show that the method works on both metric-based and meta-learning approaches by testing it on top of 3 popularly known works - Prototypical Networks, Matching Networks and the MAML framework. Our method performs better than standard augmentations, which raises the question as to which augmentations are actually useful in learning robust features, which is an interesting avenue for future work.

References

[1] alirezazareian. Issue #5: Reproducing Mini-Imagenet Results. https://github.com/jakesnell/prototypical-networks/issues/5, 2018. 3

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 2

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 1

[4] Thomas Boquet, Laure Delisle, Denis Kochetkov, Nathan Schucher, Boris N. Oreshkin, and Julien Cornebise. Reproducibility and stability analysis in metric-based few-shot learning. In *RML@ICLR*, 2019. 3

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019. 2

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 2, 3, 4

[7] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. *arXiv preprint arXiv:1709.04905*, 2017. 2

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 1

[9] Oscar Knagg. Repository for few-shot learning machine learning projects. https://github.com/oscarknagg/few-shot/, 2018. 3

[10] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011. 1

[11] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, 2011. 3, 4

[12] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1

[13] Yanbin Liu. Issue #2: Can you release detailed configuration? https://github.com/jakesnell/prototypical-networks/issues/2, 2018. 3

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[15] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 1

[16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 1

[17] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, pages 5276–5286, 2019. 1

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3, 4

[20] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 1

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[22] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1, 2, 3, 4

[23] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 1, 2

[24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 1

[25] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018. 1

[26] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1, 2, 3, 4

[27] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018. 2