MA³: Model Agnostic Adversarial Augmentation for Few Shot learning

Rohit Jena Shirsendu Sukanta Halder Katia Sycara
rjena@cs.cmu.edu shirsenh@cs.cmu.edu 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 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
to all tasks. ProtoNets [22] learn a representation based on 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 [27] 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 translation, rotation, or shear 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 $(x, y, z)^T$ and its 2D projection into the image plane $(u, v, 1)^T$. 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 ($\gamma$), yaw ($\alpha$) and pitch ($\beta$) where $||\gamma||, ||\alpha||, ||\beta|| \ll 1$, and a small change in translation $t$ such that $||t|| \ll 1$. Plugging these formulae into the rotation matrix and using Taylor expansion (ignoring third order terms and higher), we have:

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

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:

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

Since we assume it to be a distant object, and the values of $\alpha, \beta, \gamma$ are relatively small, the denominator can be simplified using binomial expansion

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

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

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

and

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

Substituting the values of $u_1, v_1$ we get

$$\begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 + \delta_1 \\ 1 + \delta_5 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}$$

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 $\delta_i$ can be determined using an adversary that detects the distortions that the model hasn’t generalized to. This is the core idea
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 classifier. However, we use it in an adversarial manner.

The optimization problem becomes:

The modified optimization problem becomes:

where \( \lambda \) 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 \( \delta \) 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 px, py and scale s and the affine matrix is constructed as:

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

The values are bounded to $\theta \in [-\theta_0, \theta_0], s \in [1 - \epsilon_s, 1 + \epsilon_s]$ and $px, py \in [-T, T]$ 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 Omniglot [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.

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

<table>
<thead>
<tr>
<th>Classification Task</th>
<th>Baseline</th>
<th>Baseline (with standard aug.)</th>
<th>Ours (λ = 0)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 way, 5 shot</td>
<td>66.6%</td>
<td>70.2%</td>
<td>58.8%</td>
<td>70.4%</td>
</tr>
<tr>
<td>5 way, 1 shot</td>
<td>51.4%</td>
<td>49.8%</td>
<td>36.2%</td>
<td>52.8%</td>
</tr>
<tr>
<td>5 way, 5 shot</td>
<td>65.9%</td>
<td>66.3%</td>
<td>57.9%</td>
<td>67.0%</td>
</tr>
<tr>
<td>5 way, 1 shot</td>
<td>47.3%</td>
<td>47.3%</td>
<td>32.1%</td>
<td>48.2%</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we introduced MA³, 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


