Skip Residual Pairwise Networks with Learnable Comparative Functions for Few-shot Learning

Akshay Mehrotra Ambedkar Dukkipati
Department of Computer Science and Automation
Indian Institute of Science
akshay13.ml@gmail.com, ambedkar@iisc.ac.in

Abstract

In this work we consider the ubiquitous Siamese network architecture and hypothesize that having an end-to-end learnable comparative function instead of an arbitrarily fixed one used commonly in practice (such as dot product) would allow the network to learn a final representation more suited to the task at hand and generalize better with very small quantities of data. Based on this we propose Skip Residual Pairwise Networks (SRPN) for few-shot learning based on residual Siamese networks. We validate our hypothesis by evaluating the proposed model for few-shot learning on Omniglot and mini-Imagenet datasets. Our model outperforms the residual Siamese design of equal depth and parameters. We also show that our model is competitive with state-of-the-art meta-learning based methods for few-shot learning on the challenging mini-Imagenet dataset whilst being a much simpler design, obtaining 54.4% accuracy on the five-way few-shot learning task with only a single example per class and over 70% accuracy with five examples per class. We further observe that the network weights in our model are much smaller compared to an equivalent residual Siamese Network under similar regularization, thus validating our hypothesis that our model design allows for better generalization. We also observe that our asymmetric, non-metric SRPN design automatically learns to approximate natural metric learning priors such as a symmetry and the triangle inequality.

1. Introduction

Human intelligence is considered the epitome towards which man-made intelligent systems strive. Two specific abilities that set the human mind apart from machines are its ability to generalize to related but unseen data and its ability to learn from small quantities of data. As machine learning methods have seen tremendous progress in recent years there has been growing focus on measuring the ability of models to generalize to unseen varieties of data and to learn from even very little of it. This is the aim of the one shot or few shot learning problem [14]. In particular, given a dataset, only a subset of whose classes are used for training, the model is tested in its ability to classify over the remaining classes using a very few examples from each class. In the case of one-shot it is just one example, in the case of few-shot it is typically 1 to 5 examples.

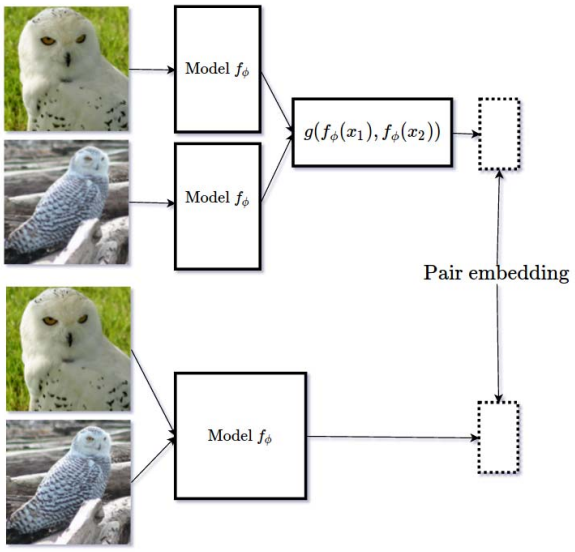

Figure 1. Two approaches to obtain the similarity embedding. In the first case (typically used in f_ϕ can be a neural network and g is an arbitrary, simple function like absolute difference or dot product. In the second case g is also parameterized and incorporated into the network f_ϕ hence being adaptive to minimize the training objective with end-to-end optimization.

End-to-end training of deep neural networks that achieves state-of-the-art results in various computer vision tasks may not suit for problems such as few-shot learning. The reason for this is deep neural networks involves learning huge number of parameters and considering that there is very little class specific data, the models may over fit.

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00099

Some recent works focus on using features learned by deep neural networks for more specialized processing suited to the few-shot task [29, 20]. There is however a lack of understanding of the generalization of deep neural networks to the few-shot setting. Deep neural networks represent a class of models that are trained very differently from how humans learn. They require multiple iterations over huge datasets and millions of gradient descent updates. Yet, these networks are able to perform tasks such as image classification or object detection at par or even better than humans. Do deep neural networks also naturally generalize to related but unseen classes of data?

A very successful deep neural network model that is proposed for one-shot or few-shot learning is deep Siamese networks [12]. Siamese networks [1] were proposed for comparing the similarity of two data-points by passing each point through the network, and then applying a comparison function (such as the squared difference or the dot product operator) on the intermediate representation to provide the final output. This paradigm has been successfully used for a variety of tasks, like face recognition [23], detecting duplicate sentences [18] few-shot learning [12].

In this paper we propose Siamese networks with learnable comparison function, in an end-to-end framework, rather than being arbitrarily fixed (see Figure 1). In particular, our proposed model enhances a traditional Siamese network for pairwise comparisons using skip residual connections [8] that allows the latent features of each image at every block to affect the computation of the other. We refer to this model as the Skip Residual Pairwise Network (SRPN). It takes a pair of images as input and outputs a single similarity embedding score/embedding for the pair. It is trained end-to-end for the similarity matching objective and used for the few-shot classification task by comparing the limited examples per class with the given test case, i.e., classification by pairwise comparison. Further, we also study the behavior of the proposed model by varying the configurable properties such as mixing style, depth and size to better understand the reason for superior few-shot performance.

1.1. Contributions

The main contributions of this work are as follows.

- We propose a new Siamese network design for few-shot learning that has greater representational capacity due to the learnable comparison function and a specialized residual design that allows for adaptive comparison of latent features.
- We present an evaluation and analysis of proposed model and its variations in mixing style, depth and size, where we focus on the properties of the model that make it better at few-shot learning than equivalent Siamese models.
- We show that our proposed model achieves state-of-the-art performance on the challenging mini-Imagenet dataset.
- Though the proposed model is not explicitly trained to learn a symmetric embedding, we show that the model exhibits convergence towards symmetry without any explicit feedback. We consider this is one remarkable property of our model.

2. Preliminaries

2.1. Few Shot Learning

The objective of few shot learning is to learn to classify well across classes unseen during training, a few examples of which are provided at test time. The model is trained on a labeled training data, where each example comes from a subset of the total classes $C_{\text{train}} \subset C$. During the test phase, the model is provided with a single example (for the one-shot case) from each of the chosen classes in $C_{\text{test}} \subset C - C_{\text{train}}$. Only using these examples as the support set S, the task is to correctly classify a sample chosen from C_{test}. The number of classes in C_{test} (5 or 20 in our case) and the number of examples (1 or 5 in our case) characterize the problem.

2.2. Siamese Networks for Few-Shot Learning

A Siamese network compares similarity between two points x_1 and x_2 by passing both through two copies of the same network f. Then one combines the two pathways by computing the similarity using a fixed function such as absolute difference or dot product. In particular, the following contrastive loss function is minimized [7].

$$\min_{\theta} \mathbb{E}_{(x_1,x_2,y)\sim p_{\text{data}}} \mathbb{1}(y=1)\|f_\theta(x_1) - f_\theta(x_2)\| + \mathbb{1}(y=0)(\delta - \|f_\theta(x_1) - f_\theta(x_2)\|)_+ \ , \quad (1)$$

where $y = 1$ if x_1 and x_2 belongs to the same class, otherwise $y = 0$. Here, θ denotes the parameters of the network, δ is the margin for minimum separation of examples belonging to different classes, and $\|.\|$ is the underlying metric or similarity measure. When using the network for classification one feeds $\|f(x) - f(x_t)\|$ into a linear logistic classifier and the network is trained using binary cross-entropy loss. Now, Siamese networks can be used for few-shot learning by comparing the similarity of the test point x_t with examples of all classes in the support set S.

2.3. Related Work

The few shot learning problem [16] has traditionally been studied under two paradigms.

(i) In a Bayesian setting that involves inferring the parameters as a posterior distribution assuming a prior and

calculating the likelihood using the given data at test time.

(ii) In a metric learning setting where the model is trained to minimize the chosen metric on the training set and assuming the inductive bias is transferred through proper optimization or regularization

The major practical difference between the two approaches is that the Bayesian methods updates the model for each test case, but the training and prediction mechanism are unchanged. On the other hand, in a metric learning setting, one uses the model learned to directly predict on the test cases but uses a different approach (pairwise comparisons) for prediction.

Some of the earliest work on few-shot learning follows the Bayesian paradigm [5, 22]. The novel approach of [14] develops the framework of Bayesian Program Learning that naturally inculcates predicting by learning from small amounts of data as a probabilistic program. The roots of the metric learning paradigm can be traced to the work on similarity matching using Siamese networks [1] that was extended for convolutional networks for face verification in [2]. Koch et al [2015] [12] demonstrate results for convolutional neural networks on the few-shot classification tasks on the Omniglot dataset. Various formulations for the similarity objective such as contrastive loss [7] and triplet loss [30] and efficient methods to do the same [23, 9]. Recently, there have been have been attempts to use meta learning approaches to directly learn the parameters for few-shot learning [20, 6].

Some of the works that most closely related to ours includes methods that try to use deep networks for learning adaptive embeddings such as [29] and [24]. These works use the full context of the input set by feeding it into a recurrent network and matching the training and testing setting for few-shot learning by constructing 1-of-K classification settings during training. Another recent work [25] is an extension of the nearest class mean idea [15] with the explicit optimization of the few-shot learning objective by essentially training the model for multi-class classification following the idea proposed in [29] of matching the training and test setup.

Our work uses pairwise comparisons prevalent in typical similarity matching settings. Our adaptive function is also tied to the depth of the network unlike the Matching Network setting, where a final embedding is adapted by feeding it to a separate recurrent network. We aim to show that deep residual networks are naturally well suited to generalization objectives and our proposed Skip Residual Pairwise Network is a natural, scalable extension of the residual network for the task of few-shot learning through similarity matching.

3. Model

3.1. Skip Residual Pairwise Network

Siamese networks have two identical feed-forward paths f with shared weights to generate embeddings for two datapoints and the similarity is then computed as a function g of these two embeddings. While this is a successful paradigm, it can be argued that the choice of g is arbitrary and the model is forced to learn an embedding that works well with a particular choice of g. Such a choice may often be suboptimal, such as when the embedding itself is an intermediate step in a solving a problem such as similarity matching for few-shot learning. To improve upon this we propose a network which takes a pair of data points as an input and outputs a single embedding. This network is trained end-to-end to optimize a particular objective – in our case the similarity between two data points.

To best achieve a similarity comparison, the network should be ideally designed to ensure the following two objectives.

(i) adequate mixing of latent features for each image ie. allow the intermediate representation of one data point to influence the intermediate representation of the other data point, instead of mixing only at the final layer as in normal Siamese networks, and

(ii) maintain the residual structure which allows gradient propagation for very deep networks.

We achieve these objectives with the network design presented in Figure 2. We can obtain the modified network design (SRPN) from an equivalent residual Siamese network in the following way.

- The initial part of both networks serves as the feature extractor for common, low-level features - it consists of multiple residual blocks through which both data points are independently passed.
- While in a residual Siamese network the second part is similar to the first and consists of simple residual blocks, in the SRPN the two inputs are passed through separate but equivalent pathways after the initial part of the network. This is done by computing a joint representation of the pair of points at each stage and modifying the skip connections in the residual network to skip a residual block
- The final part involves output of a single fixed dimension vector in contrast to the merging of two vectors using an operator g in the residual Siamese case.

The residual blocks are consistent in both models and consist of two 3×3 convolutions with Batch Normalization [10] and a non-linearity with 1×1 convolutions used for skip connections. The final embedding of the SRPN is adaptive and the expressive power of this adaptive function is tied directly to the network depth ensuring model scalability similar to residual networks. This is

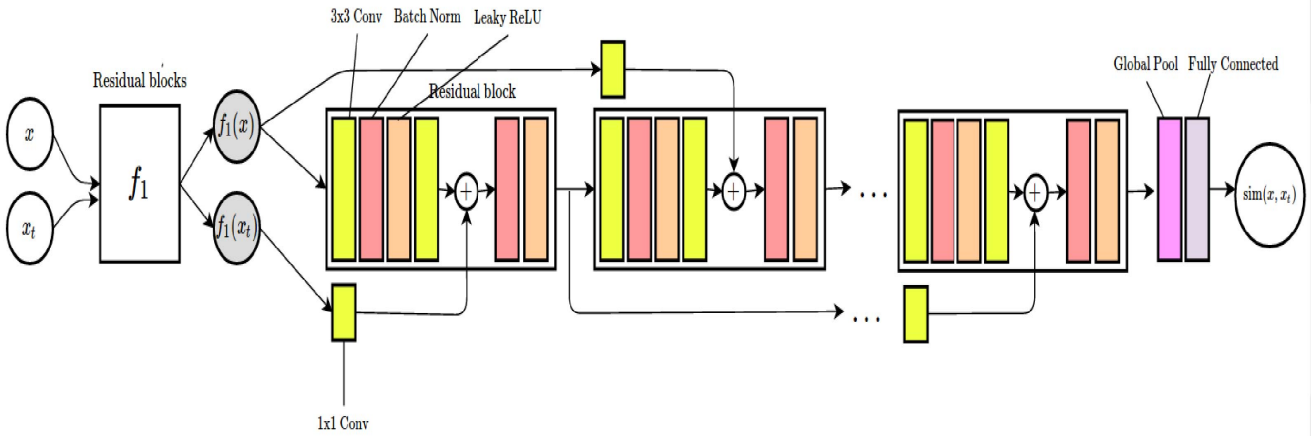

Figure 2. The proposed model Skip Residual Pairwise Net (SRPN). The network separates the intermediate computations for the inputs x and the image being compared x_t, which are then passed through equal but separate pathways using the residual connections. The final output is a single similarity vector for the pair where the distance measure is itself learned by the network.

necessary when working with complex, high-dimensional data such as real-world images. Note that the choice of the network design used for mixing can be varied, and we compare configurations with varied depth, filters per layer, depth of the initial part of shared features etc. to obtain the architecture that gives the best performance.

The final embedding output by the SRPN is fed into a linear logistic classifier with similar or dissimilar as the labels for the given pair of inputs. The whole network is trained end-to-end using the binary cross-entropy loss similar to a normal Siamese network.

3.2. Representational power

It is worth discussing at this point whether our model can learn more complex representations compared to ordinary Siamese networks. We follow the notation used above for the embedding function f, comparison function g, and the two images being compared are x and x_t. Then, the network can be seen as outputting $g(f(x), f(x_t))$. For example, in a traditional Siamese network g is the inner product or squared difference function. Assume that there exists an optimal comparison function h (by optimal here we mean a function that would best minimize the error given the input data distribution), then we would like to know whether there exists a corresponding embedding function f' such that $g(f'(x), f'(x_t)) = h(f(x), f(x_t))$ for all pairs $(x, x_t) \sim p_{data}$. We can show that such an embedding function need not exist - a simple counterexample can be constructed by assuming all variables as scalars, $f(x) = x, h(x, x_t) = x^2 + x_t^2 + x.x_t$ and $g(x, x_t) = x.x_t$ following a Siamese network with g as the dot-product operation. Thus we can argue that choosing an arbitrary, fixed comparison function g limits the modeling capacity of the combined network. We empirically demonstrate the superiority of allowing an end-to-end learnable comparison function in our model.

3.3. Symmetry of the proposed network

Note that unlike the Siamese network, the proposed SRPN does not ensure symmetric output for the model - ie. $f(x, x_t) \neq f(x_t, x)$. The model can be easily made symmetric by combining the output of $f(x, x_t)$ and $f(x_t, x)$, however we did not find any benefit from the increased computation of doing two forward passes and consider only the asymmetric design in this paper. Instead, we consider using $\|f(x, x_t) - f(x_t, x)\|_2^2$ which only requires two forward passes at training time as an additional regularizer in the loss function. However, empirically we find this is also not beneficial as the SRPN automatically moves towards a symmetric output as the training progresses. We detail this behavior in the discussion section.

3.4. Generalization with deeper networks

A general question that can be asked when training deep neural networks for the few-shot learning problem is whether performance can be improved by simply increasing the network size or number of parameters. This would require that the network's ability to learn features that generalize better should increase with network size. We found this not to be the case beyond a certain depth of the network, both with the Siamese and our proposed SRPN design. This further necessitates the improvement in network design or training methodology to improve performance in the low-

data regime.

4. Experiments

We perform experiments to validate the proposed models for few shot learning over two datasets - the simpler Omniglot and the more challenging mini-Imagenet. All our models are implemented using the Lasagne library [3] for Theano [27].

Algorithm 1 Testing protocol for N-way one shot learning

```
total-correct-pred = 0
for t = 1 to num-tests do
  sample a set C of N classes
  sample support set S consisting of 1 example from each class in C
  for r = 1 to runs do
    sample a new test point x_t from a class in C
    if predict(S, x_t) == class(x_t) then
      total-correct-pred += 1
    end if
  end for
end for
return total-correct-pred/(num-tests*runs)
```

4.1. Omniglot

Omniglot was introduced in [14] for the express purpose of measuring few-shot learning performance of models. It consists of 1623 classes of characters from 50 alphabet with 20 binary images per class. Results on the dataset have been reported using two different configurations in literature - a within alphabet setting of [14] and the more recent across alphabet setting of [29]. Within-alphabet setting involves using the initial 40 alphabet for training and validation while the remaining 10 are used for testing. The test cases are constructed by picking an alphabet from the test set and then picking N different characters from the alphabet. In contrast the across-alphabet setting uses the first 1200 classes as the training and validation set and the test cases are constructed by picking a character each from any 20 classes in the test set. In this paper we follow [29] and use the across alphabet setting to have consistent results with other recent works.

The dataset is divided into two parts - the first 1200 classes are used for training and validation, while the remaining are test classes for the few shot tasks. For this dataset we follow the evaluation protocol described in Algorithm 1 with `num-tests` = 200 and `runs` = 20.

4.1.1 Baselines

Our baselines are based on the traditional Siamese network design using AlexNet [13] style neural network and using the Wide Residual Network (WRN) [31] as the embedding functions f. Wide Residual Networks are currently one of the most effective designs for supervised vision tasks. They have two hyper-parameters: depth (d) and filter multiplicity (k). The whole network is divided into residual stacks, and each stack has k times the number of convolutional filters per layer than the previous stack. The total depth of the network is d which depends on the number of stacks and the number of residual blocks per stack. Specifically, we use the following baselines:

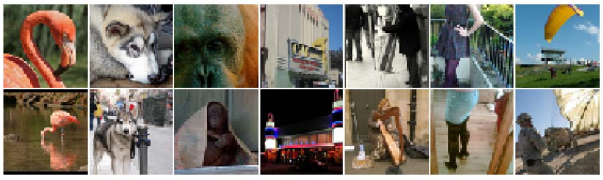

Figure 3. Actual data samples from some of the test classes for the Mini-Imagenet Dataset. Each column shows two images from the same class. Given the huge inter-class variation in these natural images, the problem of recognizing the class by comparing it to a single image of each of the other classes is challenging.

1. Siam-I: a smaller one with 5 convolutional layers and a final global pooling layer

2. Siam-II: a larger network with 5 residual blocks followed by global pooling. We follow the Wide Residual Networks design with $k = 2$

We modify Siam-II network for constructing our equivalent SRPN model. Note that while the depth of Siam-I and Siam-II is different, both have a similar number of trainable parameters.

To ensure consistency with other published results, we re-scale the images to 28x28 pixels, and augment the training data with random rotations (by ± 45 degrees and/or translations (6 pixels) in both X and Y axis. The training is done using mini-batch gradient descent with rmsprop updates and batch size of 128. The initial learning rate is set to 1×10^{-3}. L2 regularization is used with all models. We maintain a validation set of 60 classes from training for early stopping and hyperparameter validation. Results are reported in Table 1.

4.2. Mini-Imagenet

Mini-Imagenet was introduced recently by [29] as a more challenging dataset for few-shot learning tasks. The dataset consists of 100 classes of natural images from the Imagenet dataset [21] with 600 RGB images per class, rescaled to 84x84 pixels. We use the standard splits for the dataset released by [20].

The dataset is divided into three parts. The first 64 classes are used for training, the next 16 for validation,

Model	1 shot
Pixel Distance [29]	26.7
Matching Nets w/ fine-tuning [29]	93.5
Neural Statistician [4]	93.1
Convolutional ARC [24]	**97.5**
Prototypical Networks [25]	96.0
Siamese + Memory [11]	95.0
Relation Net [26]	**97.6**
Baseline (Siam-I)	88.4
Baseline (Siam-II)	92.0
SRPN	95.0

Table 1. **Omniglot :** Accuracy (%) comparison for metric learning based approaches for the 20-way one shot learning experiment on the Omniglot. Our proposed model outperforms an equivalent Siamese Network baseline (Siam-II).

while the remaining 20 are test classes for the few shot tasks. For this dataset we follow the evaluation protocol described in Algorithm 1 with `num-tests` = 100 and `runs` = 100.

4.2.1 Baselines

Since network depth has been shown to be essential for good performance on Imagenet, we restrict ourselves to the Wide Residual Network (WRN) baseline for this experiment. We choose a Siamese network with a WRN of depth (d)=40 and multiplicity (k)=2. We obtain an equivalent SRPN by modifying the WRN design.

For measuring the generalization behavior with varying depth and parameters we also test on SRPN modified from WRN with $d = 50, k = 2$ and $d = 40, k = 4$. The latter has more than twice the number of parameters compared to our original model of $d = 40, k = 2$.

To ensure consistency with other published results, we re-scale the images to 84x84 pixels. No data augmentation or pre-processing of any kind is done other than scaling in $[0, 1]$. The training is done using mini-batch gradient descent with rmsprop updates and a batch size of 64. All models are trained for a maximum of 150000 updates. We also experiment with early-stopping using the validation set error but find little difference on the test set compared with running for the fixed number of updates. The initial learning rate is set to 1×10^{-3}, and is annealed to 5×10^{-5}. L2 regularization is used with initial value of 5×10^{-7}. Hyperparameters are tuned using grid search on the validation set. Results are reported in Table 2.

Model	1 shot	5 shot
Pixel Distance [29]	23.0	26.0
Siamese WRN (k=2,d=40)	48.2 ± 0.5	63.6 ± 0.4
Matching Nets FCE [29]	43.6 ± 0.8	55.3 ± 0.7
Convolutional ARC [24]	49.1	-
Prototypical Networks [25]	49.4 ± 0.8	68.2 ± 0.7
mAP-SSVM [28]	50.3 ± 0.8	63.9 ± 0.7
Relation Net [26]	50.4 ± 0.8	65.3 ± 0.7
SRPN (k=2,d=40)	53.8 ± 0.6	68.0 ± 0.5
SRPN (k=4,d=40)	54.0 ± 0.7	68.8 ± 0.8
SRPN (k=2,d=50)	54.4 ± 0.6	$\mathbf{70.2 \pm 0.6}$
Meta-Learner LSTM [20]	43.4 ± 0.7	60.6 ± 0.7
MAML [6]	48.7 ± 1.8	63.1 ± 0.9
MetaNet [19]	49.2 ± 0.9	-
SNAIL [17]	$\mathbf{55.7 \pm 1.0}$	68.9 ± 0.9

Table 2. **mini-Imagenet :** Accuracy (%) for the 5-way few shot learning experiments on the mini-Imagenet dataset. The first section contain metric learning based approaches including our proposed SRPN model, while the second section contains meta learning based methods. The best results in each section are underlined, and best overall are in bold.

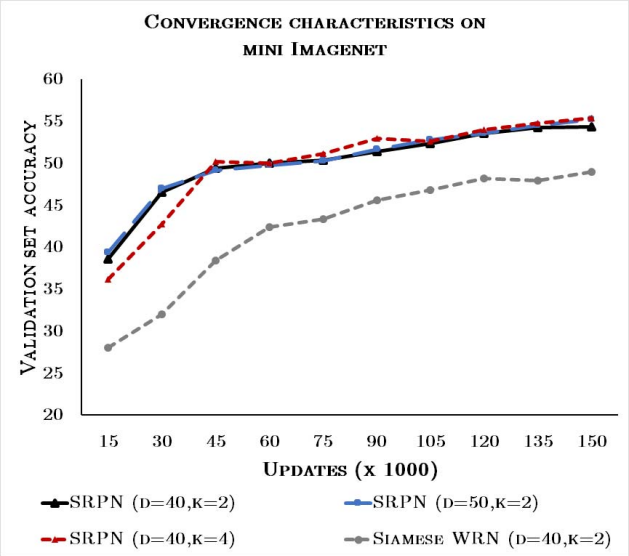

Figure 4. One-shot validation accuracy convergence characteristics of the WideResNet (WRN) Siamese baseline (k=2,d=40) versus various configurations of the Skip Residual Pairwise Networks on mini-Imagenet. Note that k refers to the multiplicity of filters and d is the depth of the network.

5. Analysis

Our first observation is that not only does the SRPN achieve higher accuracy than the equivalent WRN-based Siamese network but also shows faster convergence and the regularization loss and thus the networks weights are smaller (see Figure 4,6). In our opinion, this happens because the

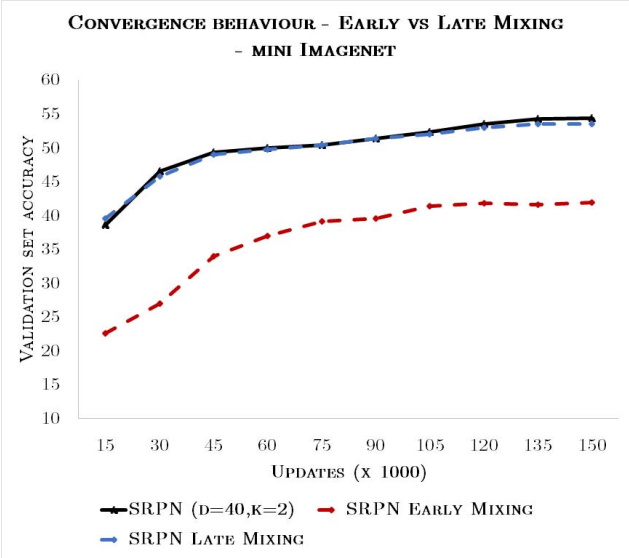

Figure 5. One-shot validation accuracy convergence characteristics based on the the mixing used in SRPN design - early mixing refers to stacking both images channel-wise and passing through a WideResNet, late mixing refers to having the shared part for both images twice as long as the original design (SRPN).

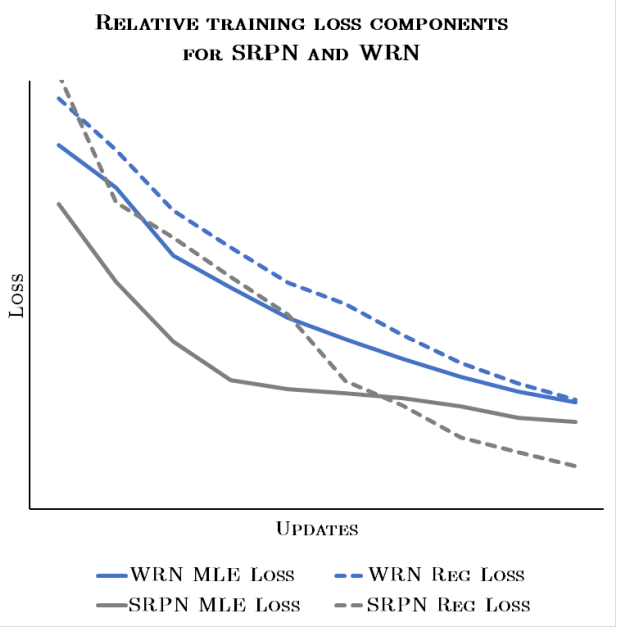

Figure 6. Loss components of the SRPN and WRN on mini-Imagenet. The SRPN has noticeably smaller network weights as the training progresses.

SRPN is not forced to learn an embedding that works well with a fixed distance metric, instead it is able to adapt to the distance measure that best minimizes the total loss. This allows the network to find a manifold which reduces the similarity loss as well as the regularization penalty, effectively leading to better generalization performance.

A remarkable observation is that while the SRPN is not explicitly trained to learn a symmetric embedding the model does this automatically as can be seen by the diminishing difference between the embeddings (pre-final layer) in Figure 5. A similar observation can be seen with regards to the violation of the triangle inequality in Figure 8.

We also note that increasing the network size does not increase performance on the few-shot learning task - the SRPN ($d = 40, k = 4$) model has more than twice the parameters of SRPN ($d = 40, k = 2$) model yet performance is very similar (Figure 4).

Also interesting to note is that for the mini-Imagenet task, our Siamese residual net baseline performs better than many proposed approaches, which reinforces the importance of depth in challenging image recognition tasks and is also an indicator that deep convolutional models with millions of parameters can be successful in learning features that generalize well to unseen data distributions and thus do well in the few-shot learning setup.

We have experiment with our core idea of an end-to-end learnable similarity network in multiple network designs. While we chose to do this using the specified architecture (Figure 2), we alternatively consider simpler methods like stacking two images along the channel axis and passing them through a vanilla residual network. We found that it led to performance significantly worse than the proposed SRPN model while also being much slower to converge (Figure 5). We argue that this shows the need for maintaining and propagating the latent representation of both images is essential in learning a joint embedding, and that by combining them at pixel level rather than at latent representation level is not useful because the pixels are much more correlated between images than within images. Having a longer common pipeline led to slightly worse results as well which indicates that mixing is essential after the most abstract features have been learned by the model.

6. Conclusion

Since the arbitrarily fixed comparison function in Siamese networks is a major limitation in similarity matching for few-shot learning, we have proposed a new model Skip Residual Pairwise Network (SPRN) with a learnable comparison function. We argue that the increased modeling capacity would allow SRPNs to be better at similarity matching in the low data regime. Our Skip Residual Pairwise Network outperforms an equivalent residual Siamese Network and achieves performance competitive with the state of the art, meta-learning models [17] while being a much

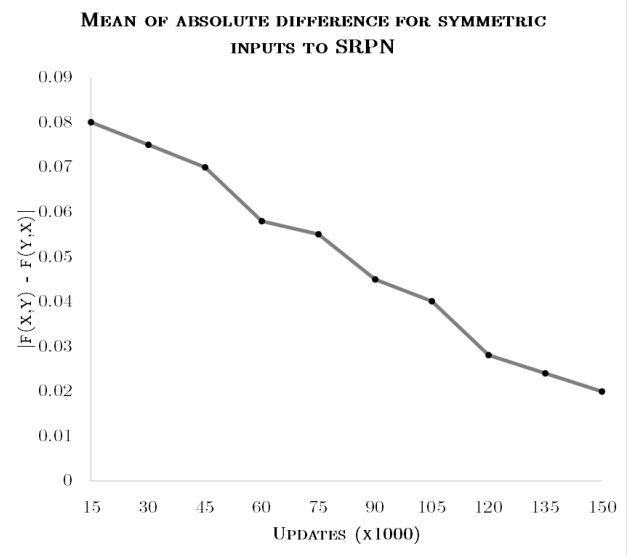

Figure 7. Mean of the difference between embeddings for symmetric inputs for the SRPN model ie. $|f(x, y) - f(y, x)|$ as training progresses, on the mini-Imagenet dataset. The model converges towards symmetric behavior without explicitly being trained to do so.

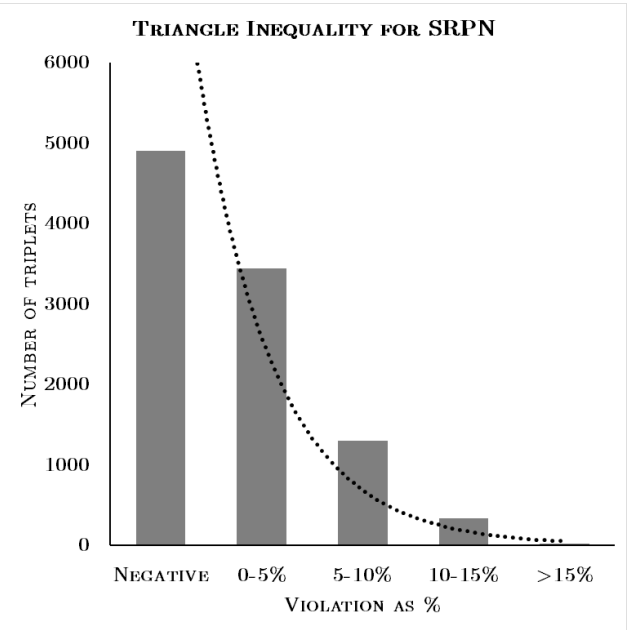

Figure 8. Violation of the triangle inequality for three randomly selected pairs of points passed through SRPN. The violation is indicated as the % of $|f(x_1) - f(x_2)| + |f(x_2) - f(x_3)|$ by which $|f(x_1) - |f(x_3)|$ exceeds it.

simpler design. One remarkable characteristic of SRPN is that it automatically learn to recover symmetry. Further, we show that network design is crucial to improving generalization to unseen classes as seen by the variations in the SRPN mixing style and that size of the network only helps till a certain extent in this regard. Future work would focus on using SRPNs in more general metric learning settings.

Acknowledgement

Authors acknowledge financial support from the CyberGut expedition project by the Robert Bosch Centre for Cyber Physical Systems at the Indian Institute of Science, Bengaluru.

References

[1] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a 'siamese' time delay neural network. *IJPRAI*, 7(4):669–688, 1993.

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546. IEEE, 2005.

[3] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, D. M. de Almeida, B. McFee, H. Weideman, G. Takcs, P. de Rivaz, J. Crall, G. Sanders, K. Rasul, C. Liu, G. French, and J. Degrave. Lasagne: First release., Aug. 2015.

[4] H. Edwards and A. Storkey. Towards a neural statistician. In *International Conference on Learning Representations (ICLR)*, 2017.

[5] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[6] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference of Machine Learning*, 2017.

[7] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742. IEEE, 2006.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[11] L. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. In *International Conference on Learning Representations (ICLR)*, 2017.

[12] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In

Advances in neural information processing systems, pages 1097–1105, 2012.

[14] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[15] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. *Computer Vision–ECCV 2012*, pages 488–501, 2012.

[16] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 464–471. IEEE, 2000.

[17] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations (ICLR)*, 2018.

[18] J. Mueller and A. Thyagarajan. Siamese recurrent architectures for learning sentence similarity. 2016.

[19] T. Munkhdalai and H. Yu. Meta networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[20] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[22] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *ICML Unsupervised and Transfer Learning*, pages 195–206, 2012.

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[24] P. Shyam, S. Gupta, and A. Dukkipati. Attentive recurrent comparators. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3173–3181, 2017.

[25] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[26] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[28] E. Triantafillou, R. Zemel, and R. Urtasun. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems*, pages 2255–2265, 2017.

[29] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[30] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[31] S. Zagoruyko and N. Komodakis. Wide residual networks. *British Machine Vision Conference (BMVC)*, 2016.