Meta-Learning with Shared Amortized Variational Inference (Supplementary Material)

A. Network Architectures

We learn separate amortized inference networks to predict the mean μ and log-variance $\ln \sigma^2$ of the latent classification weight vectors w^t. Both networks have the same architecture, which depends on the feature extractor that is used. The inference networks are shared between the prior and approximate posterior distributions.

A.1. CONV-5 Feature Extractor

The embedding of the image returned by the CONV-5 feature extractor is a 256-dimensional vector. Each of the inference networks for the mean and log variance of the classifier weights w^t consists of three fully connected layers with 256 input and output features, and ELU non-linearity (Clevert et al., 2016) between the layers. There are two additional inference networks that predict the mean and log variance of the classifier biases b^t. Both of them consist of two fully connected layers with 256 input and output features followed by ELU non-linearity, and a fully connected layer with 256 input and a single output feature. The design is the same as used by Gordon et al. (2019) to ensure comparability.

A.2. ResNet-12 Feature Extractor

With the ResNet-12 feature extractor, every image is embedded into a 512-dimensional feature vector. Each of the two inference networks consists of three fully connected layers with 512 input and output features, with skip connections and swish-1 non-linearity (Ramachandran et al., 2017) applied before addition in the first two dense layers.

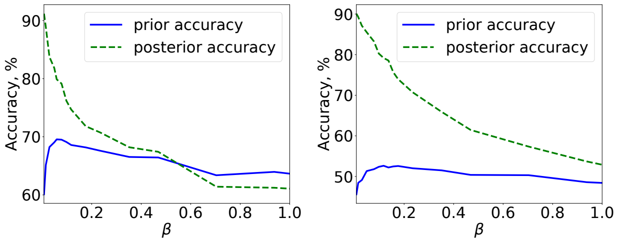
B. Training Details

For comparison with TADAM (Oreshkin et al., 2018) we use the same optimization procedure, number of SGD updates, and weight decay parameters for common parts of the architecture as in the paper. For experiments with data augmentation on miniImageNet we use 40k SGD updates with momentum 0.9, and early stopping based on meta-validation performance. We set the initial learning rate to 0.1, and decrease it by a factor ten after 20k, 25k and 30k updates. On FC100 and CIFAR-FS, we use 30k SGD updates with the same momentum and initial learning rate, and the latter is decreased after 15k, 20k and 25k updates. We clip gradients at 0.1, and set separate weight decay rates for the feature extractor, TEN, fully connected layer in the auxiliary task, and inference networks. For the feature extractor and TEN the weight decay is 0.0005. For the fully connected layer in the auxiliary task the weight decay is 0.00001 on miniImageNet, and 0.0005 on FC100 and CIFAR-FS. In the 1-shot setup, the inference networks are regularized with the weight decay equal to 0.0005, regardless of the dataset. In the 5-shot setup, the weight decay parameter in the inference networks is 0.00001 on miniImageNet, and 0.00005 on FC100 and CIFAR-FS. We empirically find that the regularization coefficient $\beta = \frac{K}{Nd}$ produces good results, and it can be used as a starting point for further parameter tuning. Here d is the dimensionality of the feature vector f_θ, N is the number of classes in the task, and K is the total number of query samples in the task. On CONV-5, we set β to 0.0586 for the 5-shot setup, and we multiply it by two for the 1-shot setup. On ResNet-12, we set β to 0.0125 for both setups, and we use a value of β twice as large for the 1-shot setup without auxiliary co-training.

For the 5-shot setup, mini-batches consist of two episodes, each with 32 query images. For the 1-shot setup, we sample 5 episodes per mini-batch, and 12 query images per episode. In both cases query images are sampled uniformly across classes, without any restriction on the number per class. The auxiliary 64-way classification task is trained with the batch size 64.

C. Impact of β-scaling

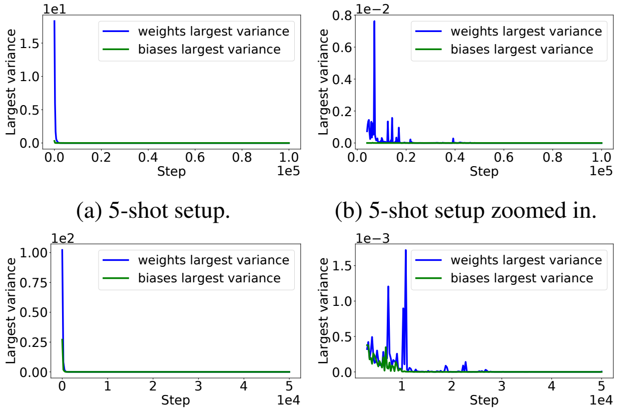
Typically, in autoencoders the dimensionality of the latent space is smaller than of the observed. This is not the case in the meta learning classification task where the output is merely a one-hot-encoded label of the class, while the latent space is of the same size as the output of the feature extractor. In our experiments we observe that the large KL term suppresses the reconstruction term resulting in a weaker performance. In particular, there is a trade off between these parts of the objective function $\hat{\mathcal{L}}(\Theta)$ which can be regulated by β-scaling of the KL term. Figure 1 shows the accuracy of SAMOVAR-base with CONV-5 feature extractor as a function of β. Even though in both setups there is a clear maximum, overall, the model is relatively robust to the setting of β. Let's denote the optimum β as β_{opt}. Then for the 5-shot setup the range at least from $0.83\beta_{\text{opt}}$ to $2\beta_{\text{opt}}$ produces results that are within the 1% interval from the



(a) 5-shot setup

(b) 1-shot setup

Figure 1. Mean accuracy of the SAMOVAR-base classifiers sampled from the prior and posterior as a function of β. While training, we fix the random seed of the data to generate the same series of miniImageNet tasks. The evaluation is performed over 5000 random tasks.



(a) 5-shot setup.

(b) 5-shot setup zoomed in.

(c) 1-shot setup.

(d) 1-shot setup zoomed in.

Figure 2. Largest variance in VERSA as a function of the optimization step. Results for optimization steps from Figure 2a and Figure 2c that follow the first encounter of variance below 0.001 are zoomed in Figure 2b Figure 2d respectively.

maximum accuracy at β_{opt}. For the 1-shot setup, the same holds true for the range at least from $0.66\beta_{\text{opt}}$ to $2\beta_{\text{opt}}$.

D. Posterior Collapse in VERSA

While training VERSA, every 250 optimization steps we keep track of the largest variance of the weights and biases of the predicted classifier. Figure 2 shows how this variance decreases with time. For example, the largest variance of the weights first falls below 0.001 at the step 4000 in the 5-shot setup, and at the step 3000 in the 1-shot setup.

References

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*, 2016.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. Meta-learning probabilistic inference for prediction. In *ICLR*, 2019.

Oreshkin, B., López, P. R., and Lacoste, A. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.