The Neural Statistician: One-Shot Handwritten Character Recognition

TJ Bai
Johns Hopkins University
tbai4@jhu.edu

Arijit Nukala
Johns Hopkins University
anukala1@jhu.edu

Abstract

We investigate the “neural statistician" (Edwards and Storkey, 2017) as an extension of the variational autoencoder for generative modeling at the *class* (dataset) level. We pose that jointly modeling sparse datasets enables data- and parameter-efficient few-shot learning and we empirically confirm these results with strong few-shot classification accuracy on the Omniglot (Lake et al., 2015) family of datasets. We further demonstrate that accuracy is robust at various training set sizes, but still lags behind human accuracy and leading meta-learning methods such as BPL.

1 Introduction

Human learners have the powerful and innate ability to generalize new concepts from relatively few examples, a phenomenon known as few-shot learning. In contrast, machine learners often require extensive training datasets and/or high-levels of supervision. This raises an interesting question bridging cognitive science and machine learning: what inductive bias can we introduce to facilitate sample-efficient few-shot learning in a machine?

Classically, learning can be posed as the problem of producing a density estimator $\hat{p}$ given a collection of samples $\{(x_i, y_i)\}_{i=1}^{N}$, under the assumption that the samples are drawn independently and identically distributed from the target distribution p. In learning theory, it is a well-established result that with a sufficient learning algorithm adhering to the inductive principle of empirical risk minimization, $\hat{p} \to p$ as $N \to \infty$, roughly as a corollary of Law of Large Numbers. Indeed, numerous deep generative models achieve great results under this paradigm, as discussed in §4

However, a clear limitation of these methods is that bounds on performance are highly dependent on the size and quality of the training data, both theoretically and practically. If a learner is presented with an abundance of samples generated from the same distribution, it is perhaps unsurprising that it learns a strong representation. In contrast, consider the more difficult learning problem where we aim to effectively learn multiple target distributions $p_1, \dots, p_D$ with a relatively small number of training examples per class. Moreover, at test-time we need to be capable of *transferring* generalizations to examples from *unseen* distributions.

This setting could be viewed as a more faithful representation of human experience, wherein the number of classes is on the order of the number of total samples, yet we are still able to construct reliable abstract representations and apply them to tasks such as few-shot classification, i.e. given a few examples, what else belongs to the same class? We propose that by modeling these diverse classes under a *shared* generative model, we gain statistical power and can achieve effective few-shot learning while maintaining data and parameter efficiency. Intuitively, as humans, we might expect that our concept of different objects, like apples and oranges, might exhibit substantial similarities in the conceptual latent space and thus assist in our recognition of lemons.

Preprint. Under review.

2 The Neural Statistician

Introduced by Edwards and Storkey (2017), the “neural statistician" provides a joint generative model at the *class* (dataset) level, as opposed to the *sample* (data point) level.

At its core, the neural statistician assumes that each class is generated by a shared *context*, $p(c)$, from which a series of latent variables, $p(z_N|c), p(z_{N-1}|z_N, c), \ldots$, are drawn to generate the observed data points $p(x|z_N, z_{N-1}, \ldots)$. The goal is to approximate the *statistic network*, denoted as $q(c|D)$, which provides a class-level latent representation for a given set of samples $D = \{x_1, \ldots, x_N\}$.

Notice that this approach is similar to the simple hierarchical model underlying Variational Autoencoders (VAEs) (Kingma and Welling, 2014), but with a shared context upon which each latent variable is conditioned, as illustrated in Figure 1. The addition of a “ladder" of sample-level latent variables allows the model to generalize to more complex internal structures. Thus, as with VAEs, the training objective and process are conceptually the same: to formulate an evidence lower-bound (ELBO) on the joint data likelihood in terms of both inferential and generative modules.

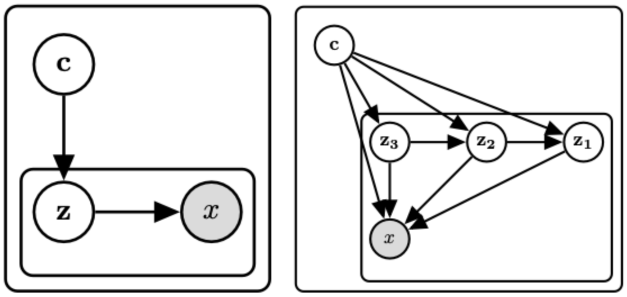

Figure 1: *Left*: VAE-like hierarchical model, *Right*: Full model, *Source*: Edwards and Storkey (2017)

2.1 Model Components

The full derivation details can be found in the original work, but here we summarize each learned module and how they relate in the ELBO formulation. Related implementation can be found at github.com/tjbai/omniglot:

- **Observation encoder,** $p(h|x)$**:** Standard 9-layer convolutional encoder that maps single-channel $28x28$ image inputs to $256x4x4$ feature maps.
- **Observation decoder,** $p(x|h)$**:** De-convolutional decoder to map embedding space back into $28x28$ input space parameterized by independent bernoulli’s.
- **Latent decoder,** $p(z_i|z_{i+1}, h)$**:** 3-layer MLP with residual connections (He et al., 2016).
- **Statistic network,** $q(c|D)$**:** The crucial component for few-shot classification. In order to capture the “exchangeable" nature of i.i.d samples from a shared class, first computes an element-wise mean across all input dimensions followed by a fully-connected layer. This netowrk outputs mean and variance parameters for a diagonal gaussian parameter space ($d = 512$) and has a spherical (unit diagonal) gaussian regularizing distribution, $p(c)$.
- **Inference network,** $q(z_i|z_{i+1}, c, h)$**:** Posteriors for latent “ladder" with same architecture as latent decoder. Regularizing distribution is corresponding decoder $p(z_i|z_{i+1}, c)$.

With each of these components, we can write the full ELBO comprised of a reconstruction loss, context divergence, and latent divergence component. Intuitively, this formulation encourages the model to accurately reconstruct the observed datasets while regularizing the latent space, as is standard with VAEs.

$$\mathcal{L}_D = R_D - C_D - L_D \tag{1}$$

$$R_D = \mathbb{E}_{q(c|D)} \sum_{x \in D} \mathbb{E}_{q(z_{1:L}|c,x)} \log p(x|z_{1:L}, c) \tag{2}$$

$$C_D = D_{KL}(q(c|D;\phi)||p(c)) \tag{3}$$

$$L_D = \mathbb{E}_{q(c,z_{1:L}|D)} \sum_{x \in D} D_{KL}(q(z_L|c,x)||p(z_L|c)) \tag{4}$$
$$+\mathbb{E}_{q(c,z_{1:L}|D)} \sum_{i=1}^{L-1} D_{KL}(q(z_i|z_{i+1},c,x;\phi)||p(z_i|z_{i+1},c;\theta))$$

2.2 Implementation Details

For completeness, we also briefly describe various relevant implementation details:

- **Class dropout:** As described in the original paper, we randomly mask out a subset of samples for each class at training time with $p = 0.2$ to prevent overfitting/improve generalization. The number of dropped samples is appended to the class-level embedding as an additional feature.
- **Loss interpolation:** We dynamically weight the reconstruction and divergence components of the loss based on epoch to emphasize the reconstruction component early on. Written formally, $\mathcal{L}_{\mathcal{D}} = (1 + 0.5^e)R_D - \left(\frac{C_D + L_D}{1 + 0.5^e}\right)$, where e is the current epoch.
- **Gradient descent:** We use the standard AdamW optimizer with initial learning rate 1e-3 and implement gradient clipping to the range $[-0.5, 0.5]$.

Unlike the original paper, we do *not* implement data augmentations, such as dilation and rotation, to synthetically increase the size and diversity of the training set due to compute limitations, so we leave this aspect to future work.

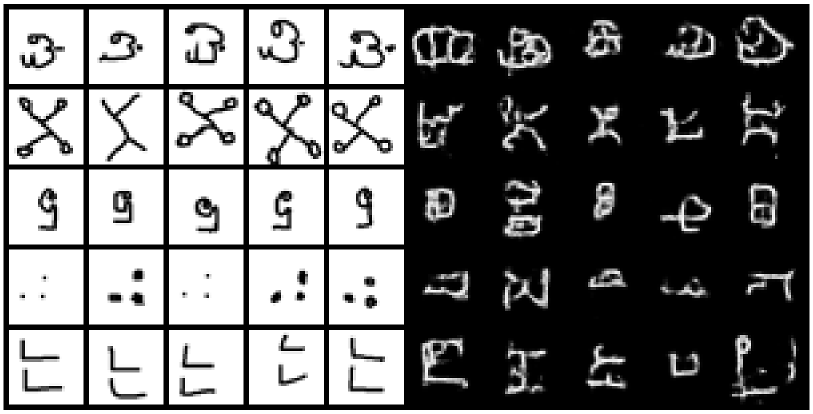

Figure 2: Sample generations (right) from 5-shot examples (left)

3 Experimental Results

We evaluate the performance of the neural statistician on the Omniglot dataset, introduced by Lake et al. (2015), specifically on the task of N-shot, K-way classification. Omniglot consists of 50 diverse language alphabets with a total of 1,623 unique characters, making it an excellent test bed for few-shot learning when we restrict training to a subset of the provided languages. We consider each character as our class-level input and a $28x28$ gray-scale representation of that character as our sample-level input. In general, each class-level input contains only 5 samples. The "large" neural statistician model is trained on 1,200 classes for 300 epochs, while the "small" model is trained on 400 classes for 1,000 epochs. To evaluate whether the class-level representation actually improves few-shot learning as hypothesized, we also train a "baseline" model where we separate 1,200 classes

	Large		Small		Paper		Baseline	
	5-way	20-way	5-way	20-way	5-way	20-way	5-way	20-way
1-shot	87.8	75.5	76.0	80.5	**98.1**	**93.2**	62.3	40.1
3-shot	93.7	86.6	90.3	79.1	-	-	-	-
5-shot	94.9	87.5	92.4	80.5	**99.5**	**98.1**	-	-

Table 1: N-shot, K-way classification results

with 5 examples each into 6,000 individual classes and train for 300 epochs. Evaluation accuracy is computed on 1,000 randomly sampled tasks from the test set. In each K-way classification task, there is one true sample from the same class and $K-1$ distractors. The predicted assignment is determined by minimizing the KL-divergence between the contexts of the test samples and the few-shot samples. Accuracy is reported as the percentage of times the model predicts the correct candidate. We only collect one-shot results for the baseline because, without a class-level representation, we do not have a formal definition for $q(c|D)$ when $|D| > 1$.

From the results collected in Table 1, we see that both large and small models obtain strong accuracies that are especially impressive in comparison to the baseline. This provides empirical confirmation for the hypothesis that jointly modeling at the class-level improves performance in the limited training data regime. As is expected, performance tends to increase as we increase N and is almost always better for $K = 5$ compared to $K = 20$. Interestingly, we see that the small model actually beats the large model specifically for $N = 1$ and $K = 20$, although this could be explained away by the larger number of training epochs.

While the results are reasonable, it is important to note that they still significantly lag behind the results reported in the paper. Additionally, other state of the art methods such as BPL achieve less than 5% error in the $N = 1, K = 20$ case and our trained models even do worse than the human baseline of roughly 8% error. This gap could perhaps be reduced if we were to implement data augmentation as in the original paper. We also note that our implementation restricts every training class to just 5 samples, whereas both BPL and the original paper make full use of the training data (the latter divides each training class with > 5 samples into many sub-classes).

4 Related Work

4.1 Few-Shot Learning

The few-shot learning paradigm centers around developing models that can generalize from few examples. Classic approaches to K-way classification include matching networks (Vinyals et al., 2016) and prototypical networks (Snell et al., 2017), which achieve competitive results to the neural statistician. However, these methods are generally discriminative and thus restricted to a specific task-based context. In contrast, the neural statistician is a full generative model.

4.2 Meta-learning

Meta-learning techniques also seek to acquire and transfer meta-knowledge from diverse contexts so that they can be applied in few-shot problems. Probabilistic approaches such as BPL (Lake et al., 2015) pose few-shot learning as a generation process and achieve state-of-the-art performance on the Omniglot classification problem, albeit through a highly task-specific model family and without benefitting from efficient amortized inference as with the neural statistician. Other methods such as MAML (Finn et al., 2017) focus on finding strong initial parameters for further gradient descent-based optimization.

4.3 Deep Generative Modeling

Latent variable models such as VAEs, in combination with deep neural networks, achieve impressive results across generative modeling. Some more recent methods such as Hierarchial VAEs (Sø nderby et al., 2016) introduce latent variables shared across samples, similar to the neural statistician, but do not provide an immediate class-based encoder that naturally transfers to few-shot learning. Other popular generative modeling paradigms that have especially succeeded in continuous domains, such as as images, include GANs (Goodfellow et al., 2014) and diffusion models (Ho et al., 2020).

5 Conclusion

We investigate the neural statistician model as a way to enable few-shot learning by modeling at the class-level. Our empirical results on the Omniglot dataset demonstrate that this approach achieves strong few-shot classification accuracy, significantly outperforming a baseline that treats each training sample individually. Although performance lags behind state-of-the-art, this deep generative model remains a promising direction for data-efficient machine learning inspired by human cognition.

References

Harrison Edwards and Amos Storkey. 2017. Towards a neural statistician. In *International Conference on Learning Representations*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

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

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Casper Kaae Sø nderby, Tapani Raiko, Lars Maalø e, Søren Kaae Sø nderby, and Ole Winther. 2016. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.