SCHA-VAE: Hierarchical Context Aggregation for Few-Shot Generation

Giorgio Giannone¹ Ole Winther¹ ²

Abstract

A few-shot generative model should be able to generate data from a novel distribution by only observing a limited set of examples. In few-shot learning the model is trained on data from many sets from distributions sharing some underlying properties such as sets of characters from different alphabets or objects from different categories. We extend current latent variable models for sets to a fully hierarchical approach with an attention-based point to set-level aggregation and call our method SCHA-VAE for Set-Context-Hierarchical-Aggregation Variational Autoencoder. We explore likelihood-based model comparison, iterative data sampling, and adaptation-free out-of-distribution generalization. Our results show that the hierarchical formulation better captures the intrinsic variability within the sets in the small data regime. This work generalizes deep latent variable approaches to few-shot learning, taking a step toward large-scale few-shot generation with a formulation that readily works with current state-of-the-art deep generative models.

1. Introduction

Humans are exceptional few-shot learners able to grasp concepts and function of objects never encountered before (Lake et al., 2011). This is because we build internal models of the world so we can combine our prior knowledge about object appearance and function to make well educated inferences from very little data (Tenenbaum, 1999; Lake et al., 2017; Ullman & Tenenbaum, 2020). In contrast, traditional machine learning systems have to be trained tabula rasa and therefore need orders of magnitude more data. In the landmark paper on modern few-shot learning Lake et al. (2015) demonstrated that with a strong prior hand-written symbols from different alphabets can be generated few-shot and distinguished one-shot, i.e. when a letter is shown for the first time.

Few-shot learning (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017) and related approaches aiming at learning from little labelled data at test time (Schaul & Schmidhuber, 2010; Hospedales et al., 2020) have recently gained new interest thanks to modeling advances, availability of large diverse datasets and computational resources. Building efficient learning systems that can adapt at inference time is a prerequisite to deploy such systems in realistic settings. Much attention has been devoted to supervised few-shot learning. The problem is typically cast in terms of an adaptive conditional task, where a small support set is used to condition explicitly (Garnelo et al., 2018a) or implicitly (Finn et al., 2017) a learner, with the goal to perform well on a query set. The high-level idea is to train the model with a large number of small sets, and inject in the model the capacity to adapt to new concepts from few-samples at test time.

![Figure 1: Lower bounds for NS-style models trained with input set size 5 varying the set cardinality from 2 to 20 on Omniglot test set. The convolutional latent space (CNS) is fundamental for performance improvement, increasing the expressivity of the context representation. Hierarchical inference over c (SCHA-VAE) and learnable aggregation (LAG) improve the generative performance in a monotonic way, providing evidence that SCHA-VAE can effectively adapt to different input set size outperforming the NS baselines.](image)

Comparatively less work has been developed on few-shot adaptation in generative models (Edwards & Storkey, 2016; Reed et al., 2017; Bartunov & Vetrov, 2018). This is partially because of the challenging nature of learning joint
latent variable models for sets through hierarchical inference and a learnable aggregation mechanism.

- We explore forward and iterative sampling strategies for the marginal and the predictive distributions implicitly defined by few-shot generative models.
- We study the few-shot transfer for this new class of models, exploring generalization to set cardinality, new classes and new datasets, providing evidence that a hierarchical set representation increases the expressivity of few-shot generative models.

2. Generative Models over Sets

In this section we present the modeling background for the proposed few-shot generative models. The Neural Statistician (NS, (Edwards & Storkey, 2016)) is a latent variable model for few-shot learning. Based on this model, many other approaches have been developed (Garnelo et al., 2018b; Bartunov & Vetrov, 2018; Hewitt et al., 2018). The NS is a hierarchical model where two collections of latent variable are learned: a task-specific summary statistic $c$ with prior $p(c)$ and a per-sample latent variable $z$ with prior $p(z|c)$:

$$p(X) = \int p(c) \prod_{s=1}^{S} \left[ \int p(x_s|z_s,c)p(z_s|c)dz_s \right] dc,$$

where $X = \{x_1, \ldots, x_S\}$ assuming the data set size is $S$ and $p(z|c)$ in general is a hierarchy of latent variables (Sønderby et al., 2016; Maaløe et al., 2019): $p(z|c) = p(z_L|c) \prod_{l=1}^{L-1} p(z_l|z_{l+1}, c)$ with $z = \{z_1, \ldots, z_L\}$.

In the original NS model, the authors factorize the lower-bound wrt $c$ and $z$ as:

$$q(c, Z|X) = q(c|X) \prod_{s=1}^{S} q(z_s|c, x_s),$$

where $Z = \{z_1, \ldots, z_S\}$ and $q(z_s|x_s, c)$ is in general a hierarchy of latent variables. In the NS, the moments of the conditioning distribution over $c$ are computed using a simple sum/average based formulation $r = \sum_{s=1}^{S} h(x_s)$; and then $r$ is used to condition $q(c|r)$. This idea is simple and straightforward, is permutation invariant, enabling efficient learning and a clean lower-bound factorization. But it has a limitation: with such formulation the model expressivity and capacity to extract information from the set are limited. The pooling mechanism assumes strong homogeneity in the context set and the generative process. However, the model formulation opens the door for more advanced invariant aggregations based on attention (Vaswani et al., 2017) and graph approaches (Velickovic et al., 2017).

In the next section, we move beyond simple aggregation to learn more expressive few-shot generative models able...
to deal with variety and complexity in the conditioning set. We propose a hierarchical merging of information between conditioning c and sample z; and a learnable aggregation mechanism for the context set.

3. Hierarchical Few-Shot Generative Models

Our goal is to learn a generative model over sets, i.e. unordered collections of observations, able to generalize to new datasets given few samples. In this model, c is a collection of latent variables that represent a set X. We learn a posterior over z conditioned on c, able to generate samples accordingly to elements x in the input set. The model has to be expressive: I) hierarchical - to increase the functions that the model can represent and improve the joint merging of set-level information c and sample information z, and II) flexible - to handle input sets of any size and complexity, improving the way the model extracts and organizes information in the conditioning set. A fundamental difference between our proposal and previous models is the intrinsically hierarchical inference procedure over c and z.

Generative Model. The generative model factorizes the joint distribution \( p(X, Z, c) \). In particular can be written as:

\[
p(X|Z, c)p(c) = p(Z_L|c) \prod_{l=1}^{L-1} p(Z_l, c_l|Z_{l+1}, c_{l+1})
\]

where \( Z_l = \{z_1^l, \ldots, z_{S_l}^l\} \) are latent for layer \( l \) in the model and sample \( s \) in the input set \( X = \{x_s\}_{s=1}^S; c_l \) is a latent for layer \( l \) in the model and the input set \( X \). We factorize the likelihood and prior terms over the set as:

\[
p(X|Z, c) = \prod_{s=1}^S p(x_s|z_s, c)
\]

\[
p(Z_l|c_l|Z_{l+1}, c_{l+1}) = \prod_{s=1}^S p(z_l^s|z_{l+1}^s, c_l) p(c_l|c_{l+1}, Z_{l+1})
\]

In this formulation the context \( p(c_l|c_{l+1}, Z_{l+1}) \) is a distribution of the previous context representation \( c_{l+1} \) and the previous latent representation for the samples \( Z_{l+1} \) as illustrated in Figure 3.

Approximate Inference. Learning in the model is achieved by amortized variational inference (Jordan et al., 1999; Hoffman et al., 2013). The hierarchical formulation leverages structured mean-field approximation to increase inference flexibility. The approximate posterior parameterizes a joint distribution between latent representation for context and samples. We factorize \( q(c, Z|X) \) following the generative model:

\[
q(c_l|X) \prod_{l=1}^{L-1} q(Z_l|c_l, Z_{l+1}, c_{l+1}, X)
\]

where we factorize the posterior using a top-down inference formulation (Sønderby et al., 2016), merging top-down stochastic inference with bottom-up deterministic inference from the data. We factorize the posterior terms over the set as:

\[
q(Z_l, c_l|Z_{l+1}, c_{l+1}, X) = \prod_{s=1}^S q(z_l^s|z_{l+1}^s, c_l, x_s) q(c_l|c_{l+1}, Z_{l+1}, X).
\]

Lower bound. In latent variable models (Rezende et al., 2014; Kingma & Welling, 2013) we maximize a per-sample lowerbound over a large dataset. In few-shot generative models we maximize a lower-bound over a large collection of small sets. This detail is important because, even if the dataset of sets is iid by construction, learning in the few-shot scenario relies explicitly on common structure within a small set. For example, each small set can be an unordered collection of observations for a specific class or concept, like a character or facet; and we rely on common structure between these observations using aggregation and conditioning on c. The variational lower bound for \( \log p(X) \) is obtained using the variational distribution \( q = q(c, Z|X) \):

\[
\mathbb{E}_q \left[ \sum_{s=1}^S \log p(x_s|z_s, c) \right] + \mathbb{E}_q \left[ \sum_{l=1}^{L-1} \sum_{s=1}^S \log \frac{p(z_l^s|z_{l+1}^s, c_l)}{q(z_l^s|c_l)} + \sum_{s=1}^S \log \frac{p(z_l^s|c_l, x_s)}{q(z_l^s|c_l, x_s)} \right] - \mathcal{KL}(q(c_l|X), p(c_l))
\]

The lower-bound can be split in three main components: an expected log likelihood term, divergences over \( Z \) and over \( c \). The final per-sample loss for \( T \) sets of size \( S \) then is \( \mathcal{L} = \mathbb{E}_p(x_0) l(X) = \frac{1}{N} \sum_{t=1}^T l(X_t) \), where \( l(X_t) \) is the negated lower-bound for set \( X_t \) and \( N = T S \).

Hierarchical Set Representation. The core idea in Eq. 8 is that a hierarchical representation for the input set improves few-shot generation. Indeed adding a hierarchy over \( c \) we largely increase the model capacities. I) The hierarchical formulation for \( p(c_l|c_{l+1}, Z_{l+1}) \) and \( q(c_l|c_{l+1}, Z_{l+1}, X) \) increases the variational posterior flexibility, and hierarchical inference has been effective to scale per-sample latent variable models to large domains (Maaløe et al., 2019; Vahdat & Kautz, 2020; Child, 2020). Following this reasoning,
a hierarchical representation for c should increase the model expressivity when dealing with sets. II) We can aggregate context set information at multiple scales and resolution by design, merging a generic c_1 with information extracted from previous resolutions. Doing so the model can iteratively refine the class information contained in the set, an essential capacity when dealing with few-samples from a new class. III) The hierarchical formulation gives us also a principle way to share per sample information z and per-set information c at different level of abstraction in the model, again increasing the model expressivity. IV) On a practical level, the hierarchical formulation can easily be coupled with modern large deep latent variable models (Child, 2020), without the risk of forgetting context information increasing depth or coarse granularity in the context, providing a promising direction for large scale few-shot generative architectures.

3.1. Sampling

**Algorithm 1 Iterative Sampling**

Input: set X
\[ c_L \sim q(c_L|X) \]
\[ z, c_{<L} \sim p(z, c_{<L}|c_L) \]
\[ x \sim p(x|z, c) \]
repeat
\[ \tilde{X} = [X, x] \]
\[ \tilde{Z}, c \sim q(\tilde{Z}, c|\tilde{X}) \]
\[ \tilde{X}' \sim p(\tilde{X}'|\tilde{Z}, c) \]
\[ x = x' \]
until Converged
return x

We may be interested in either sampling unconditionally \( x \sim p(x) \) or from the predictive distribution \( x \sim p(x|X) \). Unconditional sampling may be performed exactly using the generative model as illustrated in Figure 3: first sample the hierarchical prior \( z, c \sim p(z, c) \) and then sample the likelihood \( x \sim p(x|z, c) \). Conditional sampling \( x \sim p(x|X) \) can be done approximately using the variational posterior \( q(Z, c|X) \) as a replacement for the exact posterior \( p(Z, c|X) \) as outlined in Algorithm 1. In the single pass approach (adapted from the NS) a sample from:

\[
\int p(x|z, c)p(z, c_{<L}|c_L)q(c_L|X)dzdc = p(x|X) \quad (9)
\]
is generated. This approach is not ideal because c_{<L} is only modeled jointly with the latent z for the new sample while omitting the latent Z for X. We can introduce this dependence in a Markov chain approach adapted from the missing data imputation framework proposed in Appendix F of (Rezende et al., 2014). We augment X with the sample x generated in the single pass approach: \( \tilde{X} = [X, x] \). From this we can construct a distribution \( \hat{p}(\tilde{Z}, c|\tilde{X}) \), where \( \tilde{Z} = [Z, z] \) is the corresponding augmented latent. From this distribution and the likelihood we can construct a transition kernel:

\[
\hat{p}(x'|x, X) = \int \hat{p}(\tilde{X}'|\tilde{Z}, c, X)\hat{p}(\tilde{Z}, c|\tilde{X})dX'd\tilde{Z}dc.
\]

If \( \hat{p}(\tilde{Z}, c|\tilde{X}) = p(\tilde{Z}, c|\tilde{X}) \) then we can show that \( p(x|X) \) is an eigen-distribution for the transition kernel and thus sampling the transition kernel will under mild conditions converge to a sample from \( p(x|X) \). There are several ways to construct \( \hat{p}(\tilde{Z}, c|\tilde{X}) \) from the hierarchical variational and prior distributions. One possibility is shown in Algorithm 1. Alternatively, one may sample \( c_L, Z_L \sim q(c_L, Z_L|X) \), generate the remainder of the latent from the prior hierarchy and new samples \( x', x'' \) from the likelihood. If the variational is exact, these approaches are equivalent and exact.
the experimental section we report results for the different approaches.

3.2. Learnable aggregation (LAG)

**Algorithm 2 Learnable Aggregation (LAG)**

**Input:** set \( H = \{ h_s = h(x_s) \}_{s=1}^{S} \)

- base aggregation: \( r = 1/S \sum_{s=1}^{S} h_s \)
- compute query, key, value: \( q = g(r), k_s = k(h_s), v_s = v(h_s) \)
- weights: \( \alpha(q, k_s) = \sigma(\delta \circ \tau(q, k_s)) \)

**aggregate**

\[ r_{LAG} = \sum_{s=1}^{S} \alpha(q, k_s) v_s \]

**sample**

\( c \sim q(c \mid r_{LAG}) \)

A central idea in few-shot generative models is to condition the generative model with a permutation invariant representation of the input set. For a NS such operator is a hard-coded per-set statistic. This approach (Edwards & Storkey, 2016; Garnelo et al., 2018b) maps each sample in \( X \) independently using \( \{ h(x_s) \}_{s=1}^{S} \) and then aggregating \( r_L = \sum_{s=1}^{S} h(x_s) \) to generate the moments for \( q(c_L \mid r_L) \). This idea is simple and effective when using homogeneous and small sets for conditioning. Another choice is a relation network (Sung et al., 2018; Rusu et al., 2018) followed by aggregation. However, the adaptation capacities of the model are a function of how we represent the input set and a more expressive learnable aggregation mechanism can be useful. In the general scenario, a few-shot generative model should be able to extract information from any conditioning set \( X \) in terms of variety and size. In this paper we consider a multi-head attention-based learnable aggregation (LAG, Algorithm 2) inspired by (Lee et al., 2019) that can be used in each block of the hierarchy over \( c \) (Figure 4). Using LAG we can account for statistical dependencies in the set, handle variety between samples in the set and generalize better to input set dimensions. In the experimental section we provide extensive empirical analysis of how using LAG improves the generative and transfer capacity of the model. More details in Appendix A.

4. Experiments

In this section we discuss experimental setup and results for our model. In particular for all the models our interests are: I) Quantitative evaluation of few-shot generation; II) Few-shot conditional and unconditional sampling from the model; III) Transfer of generative capacities to new datasets and input set size; IV) Evidence that the hierarchical context representation is useful for the model.

**Datasets.** We train models on binarized Omniglot (Lake et al., 2015), CelebA (Liu et al., 2015) and FS-CIFAR100 (Oreshkin et al., 2018). For Omniglot we consider two scenarios: the train-test split (also known as background-evaluation split) proposed in (Lake et al., 2015): in this scenario all the characters at test time are from new alphabets. And the variant proposed in (Edwards & Storkey, 2016), where all the characters are mixed together and randomly assigned to train and test splits. We perform quantitative evaluation on Omniglot, MNIST (LeCun, 1998), DOUBLE-MNIST (Sun, 2019), TRIPLE-MNIST (Sun, 2019), CelebA and FS-CIFAR100. We resize all the binary datasets to 28x28, CIFAR100 to 32x32, and CelebA to 64x64.

**Training Details.** We follow the approach proposed in (Edwards & Storkey, 2016) for training with sets. We create a large collection of small sets, where each set contains all or some of the occurrences for a specific class or concept in a dataset: a character for Omniglot; the face of an identity for CelebA. Both datasets contain thousands of characters/identities with an average number of 20 occurrences per class. For this reason they are a natural choice for few-shot generation. Then we split the sets in train/val/test sets. Using these splits we can dynamically create sets of different dimensions, generating a new collection of training sets at each epoch during training. For training we use the episodic approach proposed in (Vinyals et al., 2016). Each input set is a homogeneous (one concept or class) collection of samples. During training the input set size is always between 2 and 20. The NS architecture is a close approximation of (Edwards & Storkey, 2016). We describe additional details in Appendix C.

**Baselines.** We use a VAE - which does not explore set information - and the Neural Statistician (NS) as baselines. We consider two main model variants. Such variants are characterized by different design choices for the conditioning mechanism: standard aggregation mechanism (MEAN pooling) and learnable ones (LAG). We also consider a Convolutional Neural Statistician (CNS) where the latent space is shaped with convolutions at a given resolution. We call our approach Set-Context-Hierarchical-Aggregation Variational Autoencoder (SCHA-VAE, pronounced shave) where an additional hierarchy over \( c \) is employed.

**Model Design.** NS-based models are difficult to scale and performance tends to plateau increasing depth. This can
be explained considering that \( c \sim q(c|X) \) has to be shared through all the network and information can easily be lost in the stochastic sampling process. We overcome these challenges and stabilize SCHA-VAE when increasing depth, we use a shared routing path between \( c \) and \( z \): we learn bias term \( b^c \) and \( b^z \) for the top layer, and merge all sampling information in these paths: \( b^c_l = b^c_{l+1} + f(X, b^c_{l+1}) + c_l \) and \( b^z_l = b^z_{l+1} + f(x, b^z_{l+1}) + z_l \), where \( z_l \) and \( c_l \) are samples from layer \( l \). We then share information through these path, helping to stabilize and preserve information through the hierarchy. We additionally remove any form of normalization (batch/layer norm) in the hierarchical formulation. This is not directly related to performance, but to challenge the model to perform pure few-shot generation at train and test time, without relying on batch statistics. Our implementation is inspired by (Child, 2020).

### 4.1. Test log-likelihood

With these experiments we test the generalization properties of the model on few-shot generation. We explore the behavior of the model increasing the input size. Evaluation of generative properties is performed using the log likelihood lower bound (ELBO) and approximating the log marginal likelihood with 1000 importance samples. In Table 1 and 2 we consider few-shot generalization on Omniglot and MNIST with disjoint classes and set size 5. In Figure 1 and 10 we show the models behavior increasing the input set cardinality at test time. All the models are trained with input sets of size 5 on Omniglot. Then at test time we vary the size of the context set between 2 and 20.

The NS is the original Neural Statistician with mean aggregation. CNS employs a convolutional set representation. The models with LAG use an attention-based learnable aggregation, that builds an adaptive aggregation mechanism for each set. For SCHA-VAE we use the same two aggregation mechanisms. SCHA-VAE improves performance in terms of ELBO and likelihood on different Omniglot test splits and transfer to MNIST without adaptation better than the baselines. The hierarchical formulation for \( c \) provides most of the improvement. Adding the learnable aggregation mechanism gives an additional boost. Increasing the set (2-20), SCHA-VAE can aggregate more and more information, learning a better model for the data under different regimes (in-distro, out-distro) and augmentation schemes. See Appendix, Table 3 for additional quantitative evaluation with few-shot generative models proposed in the literature, where we compare SCHA-VAE with autoregressive-based models for few-shot generation (Bartunov & Vetrov, 2018; Hewitt et al., 2018).

### 4.2. Sampling

In latent variable models like VAEs we can only perform unconditional sampling. In a NS we have different ways of sampling as explained in Section 3.1. In particular two main sampling approaches can be used: I) conditional sampling, where we sample the predictive distribution \( p(x|X) \) relying on the inference model. II) unconditional sampling, sampling \( p(X) \) through \( c \sim p(c) \) and then \( p(z|c) \). In Figure 11 we show (from left to right) stochastic reconstructions, input sets, conditional samples, and unconditional samples. For simple characters there is almost no difference between conditional and refinement sampling. However, when the model is challenged with a new complex character (Figure 11, top), the refinement procedure greatly improves the adaptation capacities and visual quality of the generated samples. In the rightmost column in the bottom part of Figure 11 we see fully unconditional samples. The model, given a set representation \( c \), generates consistent symbols, corroborating the assumption that the model learns a different distribution for each \( c \), greatly increasing the model flexibility and representation capacities.

### 4.3. Transfer

With these experiments we explore few-shot generation in the context of transfer learning. We use the same models we trained on Omniglot and we test on unseen classes in a different dataset. We use MNIST test set (10 classes), DOUBLE-MNIST test set (20 classes) and TRIPLE-MNIST test set (200 classes). The datasets increase in complexity and size. We expect relatively good performance on the
We claim that depth in the set representation and aggregation is important for good performance on concepts from different datasets. In Appendix B, we report out of distribution classification performance for the three approaches described in Section 3. The models are trained on Omniglot for context size 5 and tested on MNIST also for context size 5 without adapting the distributions to the MNIST data.

4.4. Hierarchy

We claim that depth in the set representation and aggregation at multiple resolutions are essential to increase model expressivity. In Fig. 7 we provide empirical evidence. We train SCHA-VAE with 12 stochastic layers on Omniglot, and with 24 layers on CelebA and CIFAR100. At inference, $c$ is bypassed for a certain number of layers, without modifying the depth for $z$, effectively decreasing the context depth, and we compute the test ELBO normalizing to one for easier comparison between datasets. If the hierarchy over $c$ is redundant, removing a certain number of layers should not impact the overall result; if instead the model is using aggregation at different resolutions, then the result should be greatly impacted. We see how removing the hierarchical information over $c$ worsen greatly the results for all the datasets, providing evidence that the set representation for few-shot generation. On the right we plot the cumulative KL for CelebA and CIFAR100, showing how each layer contains additional information.

5. Related Work

Learning from Sets. In recent years a large corpus of work studied the problem of learning from sets (Zaheer et al., 2017), and more generally learning in exchangeable deep models (De Finetti, 2020; Bloem-Reddy & Teh, 2020; 2019). These models can be formulated in a variety of ways, but they all have in common a form of permutation invariant aggregation (or pooling mechanism) over the input set. Deep Sets (Zaheer et al., 2017) formalized the framework of exchangeable models. The Neural Statistician (Edwards & Storkey, 2016) was the first model proposing to learn from sets in the variational autoencoder framework and used a simple and effective mean pooling mechanism for aggregation. The authors explored the representation capacities of such model for clustering and few-shot supervised learning. Generative Query Networks performs neural rendering (Esldami et al., 2018) where the problem of pooling views arises. The Neural process family (Garnelo et al., 2018b; Kim et al., 2019), where a set of point is used to learn a context set.

| Table 1: | Generalization on disjoint Omniglot classes trained on set size 5 for a V AE, NSs with mean/learnable aggregation (MEAN/LAG), convolutional variants (C) and for a SCHA-VAE with a hierarchy over $c$. We report minus the 1-sample lower bound (NELBO), and minus the 1k importance sample lower bound (NLL). $z_{1:L}$: hierarchical representation for sample. $c_{1:L}$: hierarchical representation for set. NS based models process the sets at one resolution without hierarchical inference over $c$. SCHA-VAE processes the sets at multiple resolutions with hierarchical inference over $c$. We see that the hierarchical set representation with aggregation at multiple resolutions is essential to improve generative performance. |
|---|---|---|---|---|---|---|---|
| | Agg | $z_{1:L}$ | $c_{1:L}$ | Omniglot-ns | Omniglot-back-eval | MNIST | Params (M) |
| VAE | — | ✓ | — | 101.5 ± 1 | 95.9 | 101.4 ± 1 | 88.4 | 124.7 ± 1 | 13.3 |
| NS | MEAN | ✓ | ✗ | 96.6 ± 1 | 92.4 | 91.9 ± 1 | 85.6 | 120.3 ± 1 | 14.7 |
| CNS | MEAN | ✓ | ✗ | 92.9 ± 1 | 89.7 | 83.4 ± 1 | 77.5 | 118.2 ± 2 | 16.9 |
| CNS | LAG | ✓ | ✗ | 92.8 ± 1 | 89.7 | 83.8 ± 1 | 77.6 | 117.6 ± 1 | 17.2 |
| SCHA-VAE (ours) | MEAN | ✓ | ✓ | 89.4 ± 1 | 85.8 | 79.4 ± 1 | 72.7 | 115.3 ± 2 | 11.9 |
| SCHA-VAE (ours) | LAG | ✓ | ✓ | 88.4 ± 3 | 85.4 | 77.9 ± 3 | 71.5 | 114.9 ± 1 | 12.8 |

| Table 2: | Generalization on Omniglot for different set size and augmentation scheme. We train one model on sets of size 5 and test on standard and augmented (translation, rotation, flipping) Omniglot data. For the in in-distro we test on known-classes; out-distro on unknown-classes. We use minus the 1-sample lower bound (NELBO) to evaluate the models. Lower is better. |
|---|---|---|---|---|---|---|---|
| | in-distro | Standard | out-distro | Augmented | in-distro | Standard | out-distro |
| 2 | 5 | 10 | 20 | 2 | 5 | 10 | 20 |
| NS | 96.8 | 91.2 | 89.6 | 88.8 | 101.5 | 96.6 | 95.1 | 94.7 | 107.8 | 102.6 | 101.1 | 100.5 | 111.1 | 106.1 | 104.6 | 103.9 |
| CNS | 97.8 | 91.7 | 90.0 | 89.2 | 98.8 | 92.8 | 91.2 | 90.6 | 107.7 | 101.7 | 100.0 | 99.4 | 108.6 | 102.7 | 100.9 | 100.1 |
| SCHA-VAE | 91.5 | 85.2 | 83.5 | 82.7 | 96.2 | 89.7 | 87.9 | 87.0 | 103.0 | 96.6 | 94.8 | 94.1 | 105.9 | 99.2 | 97.2 | 96.2 |
and solve downstream tasks like image completion and few-shot learning. Set Transformers (Lee et al., 2019) leverages attention to solve problems involving sets. PointNet (Qi et al., 2017) models point clouds as a set of points. Graph Attention Networks (Velickovic et al., 2017) aggregate information from related nodes using attention. Associate Compression Network (Graves et al., 2018) can be interpreted in this framework, where a prior for a VAE is learned using the top-knn retrieved in latent space. In this work we build on ideas and intuitions in these works, with a focus on generative models for sets. SetVAE (Kim et al., 2021) proposes a VAE for point-clouds, showing that processing the input set at multiple resolutions is a promising direction for set-based latent variable models.

**Few-Shot Generative Models.** Historically the machine learning community has focused its attention on supervised few-shot learning, solving a classification or regression task on new classes at test time given a small number of labeled examples. The problem can be tackled using metric based approaches (Vinyals et al., 2016; Snell et al., 2017; Oreshkin et al., 2018), gradient-based adaptation (Finn et al., 2017), optimization (Ravi & Larochelle, 2016). More generally, the few-shot learning task can be recast as bayesian inference in hierarchical modelling (Grant et al., 2018; Ravi & Beatson, 2018). In such models, typically parameters or representation are conditioned on the task, and conditional predictors are learned for such task. In (Xu et al., 2019) an iterative attention mechanism is used to learn a query-dependent task representation for supervised few-shot learning. Modern few-shot generation in machine learning was introduced in (Lake et al., 2015). The Neural Statistician (Edwards & Storkey, 2016) is one of the first few-shot learning models in the context of VAEs. However the authors focused on downstream tasks and not on generative modeling. The model has been improved further increasing expressivity for the conditional prior using powerful autoregressive models (Hewitt et al., 2018), a non-parametric formulation for the context (Wu et al., 2020) and exploiting supervision (Garnelo et al., 2018b). (Rezende et al., 2017) models point clouds as a set of points. Graph Attention Networks (Velickovic et al., 2017) aggregate in-examples. The problem can be tackled using metric based on new classes at test time given a small number of labeled few-shot learning, solving a classification or regression task Few-Shot Generative Models.

Historically the machine learning community has focused its attention on supervised few-shot learning, solving a classification or regression task on new classes at test time given a small number of labeled examples. The problem can be tackled using metric based on new classes at test time given a small number of labeled few-shot learning, solving a classification or regression task Few-Shot Generative Models.

Historically the machine learning community has focused its attention on supervised few-shot learning, solving a classification or regression task on new classes at test time given a small number of labeled examples. The problem can be tackled using metric based on new classes at test time given a small number of labeled few-shot learning, solving a classification or regression task Few-Shot Generative Models.

Historically the machine learning community has focused its attention on supervised few-shot learning, solving a classification or regression task on new classes at test time given a small number of labeled examples. The problem can be tackled using metric based on new classes at test time given a small number of labeled few-shot learning, solving a classification or regression task Few-Shot Generative Models.
2016) proposed a recurrent and attentive sequential generative model for one-shot learning based on (Gregor et al., 2015). Powerful autoregressive decoders and gradient-based adaptation are employed in (Reed et al., 2017) for one-shot generation. The context $c$ in this model is a deterministic variable. In GMN (Bartunov & Vetrov, 2018) a variational recurrent model learns a per-sample context-aware latent variable. Similar to our approach, GMN learns a flexible context, learning an attention based kernel that can handle generic datasets. However the context-aware representation scales quadratic with the input size, there is no separation between global and local information in latent space, and the input set is processed in an arbitrary autoregressive order, and not in a permutation invariant manner. Finally, recent large-scale autoregressive language models (Brown et al., 2020) exhibit non-trivial few-shot capacities.

6. Conclusion

Leveraging recent advances in deep latent variable models, we propose a new class of hierarchical latent variable models for few-shot generation. We ground our formulation in hierarchical inference and a learnable non-parametric aggregation. We show how simple hierarchical inference is a viable adaptation strategy. We perform extensive empirical evaluation in terms of generative metrics, sampling capacities and transfer properties. The proposed formulation is completely general and we expect there is large potential for improving performance by combining it with state of the art VAE architectures. Since benchmarks often come with grouping information, using a hierarchical formulation is a generic approach to improve generative capabilities.

Acknowledgement

We would like to thank Anders Christensen, Marco Ciccone, Andrea Dittadi, Pierluca D’Oro, Søren Hauberg, Valentin Liévin, Didrik Nielsen, Mathias Schreiner, Timon Willi, Max Wilson for insightful comments and useful discussions.

References

Bartunov, S. and Vetrov, D. Few-shot generative modelling with generative matching networks. In International Conference on Artificial Intelligence and Statistics, pp. 670–678, 2018.

Bloem-Reddy, B. and Teh, Y. W. Probabilistic symmetry and invariant neural networks. arXiv preprint arXiv:1901.06082, 2019.

Bloem-Reddy, B. and Teh, Y. W. Probabilistic symmetries and invariant neural networks. Journal of Machine Learning Research, 21(90):1–61, 2020.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Child, R. Very deep vae’s generalize autoregressive models and can outperform them on images. arXiv preprint arXiv:2011.10650, 2020.

De Finetti, B. 9. on the condition of partial exchangeability. In Studies in Inductive Logic and Probability Volume 2, pp. 193–206. University of California Press, 2020.

Edwards, H. and Storkey, A. Towards a neural statistician. arXiv preprint arXiv:1606.02185, 2016.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., et al. Neural scene representation and rendering. Science, 360(6394):1204–1210, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400, 2017.

Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., and Eslami, S. Conditional neural processes. arXiv preprint arXiv:1807.01613, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. arXiv preprint arXiv:1807.01622, 2018b.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. Recasting gradient-based meta-learning as hierarchical bayes. arXiv preprint arXiv:1801.08930, 2018.

Graves, A., Menick, J., and van den Oord, A. Associative compression networks for representation learning. CoRR, abs/1804.02476, 2018. URL http://arxiv.org/abs/1804.02476.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015.

Hewitt, L. B., Nye, M. L., Gane, A., Jaakkola, T., and Tenenbaum, J. B. The variational homoencoder: Learning to learn high capacity generative models from few examples. arXiv preprint arXiv:1807.08919, 2018.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. The Journal of Machine Learning Research, 14(1):1303–1347, 2013.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. arXiv preprint arXiv:2004.05439, 2020.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. Machine learning, 37(2):183–233, 1999.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. arXiv preprint arXiv:1901.05761, 2019.

Kim, J., Yoo, J., Lee, J., and Hong, S. Setvae: Learning hierarchical composition for generative modeling of set-structured data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15059–15068, 2021.
Notation.

- **Bold** → all layers
- **CAPITALIZED** → ALL SAMPLES

- **Bold CAPITALIZED**: all layers, ALL SAMPLES $Z = \{z^{s,l}_{1} \}_{s=1,l=1}^{S,L}$
- **Bold**: all layers, one sample $z_s = \{z^{s,l}_{l} \}_{l=1}^{L}$
- **CAPITALIZED**: all layers, one sample $Z_l = \{z^{s,i}_{l} \}_{s=1}^{S}$

Top Prior $c$: $p_\theta(c_L)$
Top Prior $z$: $p_\theta(z_L \mid c_L)$
Prior $c$: $p_\theta(c_l \mid c_{l+1}, Z_{l+1})$
Prior $z$: $p_\theta(z_l \mid z_{l+1}, c_l)$
Observation $x$: $p_\theta(x \mid z_{1:L}, c_{1:L})$

Top Posterior $c$: $q_\phi(c_L \mid X)$
Top Posterior $z$: $q_\phi(z_L \mid c_L, x)$
Posterior $c$: $q_\phi(c_l \mid c_{l+1}, Z_{l+1}, X)$
Posterior $z$: $q_\phi(z_l \mid z_{l+1}, c_l, x)$
Set Representation: $h_l$

---

**Figure 8**: Generation and Inference for a Neural Statistician (left) and a Hierarchical Few-Shot Generative Model (right)
A. Model Derivation

In this section we derive the generative model, inference model and lower-bound for a Basic Neural Statistician (bNS), a Neural Statistician (with hierarchy over \( z \), this is the model used as baseline in the paper) (NS), and a Hierarchical Few-Shot Generative Model (HFSGM) with hierarchy over \( z \) and \( c \). Doing so we can underline similarities and differences among the formulations.

Set Marginal. Our goal is to model a distribution \( p(X) = \prod_{s=1}^{S} p(x_s) \) where \( X \{ x_s \}_{s=1}^{S} \) is in general a small set ranging from 1 to 20 samples. We sample these sets from a common process. The Neural Statistician (Edwards & Storkey, 2016) introduces a global latent variable \( c \) for a set \( X \):

\[
p(X) = \int p(X, c) \, dc = \int p(c) \, p(X | c) \, dc = \int p(c) \, \prod_{s=1}^{S} p(x_s | c) \, dc. \quad (10)
\]

Then we can introduce a per-sample latent variable \( z \):

\[
p(X) = \int p(c) \left[ \int p(X, Z | c) dZ \right] dc = \int p(c) \prod_{s=1}^{S} \left[ \int p(x_s, z_s | c) dz_s \right] dc
\]

\[
= \int p(c) \prod_{s=1}^{S} \left[ \int p(x_s | z_s, c) p(z_s | c) dz_s \right] dc, \quad (11)
\]

where \( X = \{ x_s \}_{s=1}^{S} \) is a set of images, \( c \) is a latent variable for the set, \( Z = \{ z_s \}_{s=1}^{S} \) are latent variable for the samples in the set. The formula above is the basic marginal for all the NS-like model.

A.1. Generative Model

We can think of the model as composed of three components: a hierarchical prior over \( z \), a hierarchical prior over \( c \), and a render for \( x \). For each hierarchical prior, the top distribution is not autoregressive and act as an unconditional prior for \( c \) and a conditional prior for \( z \). In the following all the latent distributions are normal distributions with diagonal covariance. Given the hierarchical nature of our model, these assumptions are not restrictive, because from top to down in the hierarchy we build an expressive structured mean field approximation. The decoder is Bernoulli distributed for binary datasets. In the following equation we use \( Z = \{ z_s \}_{s=1}^{S}, Z = \{ z_s \}_{s=1}^{S}, z_s = \{ z_{s_l} \}_{l=1}^{L}, c = \{ c_l \}_{l=1}^{L} \). Each of these equation can be written per-set or per-sample. We choose to write everything in a compact format writing per-set equations.

bNS. In this setting both \( z \) and \( c \) are shallow latent variables:

\[
p(X, Z, c) = p(X | Z, c) p(c | Z) \quad (12)
\]

NS. \( z \) is a hierarchy:

\[
p(X, Z, c) = p(X | Z, c) p(c | Z) \prod_{l=1}^{L-1} p(Z_l | Z_{l+1}, c) \quad (13)
\]

HFSGM. Both \( z \) and \( c \) are a hierarchy. We increase the model flexibility.

Using such hierarchy we can:

- learn a structured mean field approximation for \( c \);
- jointly learn \( c \) and \( Z \), informing different stages of the learning process;
- incrementally improve \( c \) through layers of inference.
We learn the model using Amortized Variational Inference. In a NS-like model, inference is intrinsically hierarchical: the model encodes global set-level information in \( c \) using \( q(c|X) \).

**bNS.**

\[
q(c, Z|X) = q(c|X) \left[ q(Z|c, X) \right]
\]  

**NS.**

\[
q(c, Z|X) = q(c|X) \left[ q(Z|c, X) \right]
\]

**HFSGM.**

\[
q(c, Z|X) = q(c_L|X) \left[ q(Z|c_L, X) \prod_{l=1}^{L-1} q(Z_l|c_l|Z_{l+1}, c_{l+1}, X) \right]
\]

\[
q(Z_l, c_l|Z_{l+1}, c_{l+1}, X) = q(Z_l|Z_{l+1}, c_l, X) q(c_l|c_{l+1}, Z_{l+1}, X).
\]

### A.3. Lower-bound

We learn the model by Amortized Variational Inference similarly to recent methods. However we are in presence of two different latent variables, and we need to lower-bound wrt both.

**bNS.**

\[
\log p(X) \geq \mathbb{E}_{q(c, Z|X)} \left[ \log \frac{p(X, Z, c)}{q(c, Z|X)} \right] = \\
\mathbb{E}_q \left[ \sum_{s=1}^S \log p(x_s|z_s, c) \right] + \mathbb{E}_q \left[ \sum_{s=1}^S \log \frac{p(z_s|c)}{q(z_s|c, x_s)} \right] - \text{KL}(q(c|X), p(c)) = -\mathcal{L}(X).
\]

**NS.**

\[
\log p(X) \geq \mathbb{E}_{q(c, Z|X)} \left[ \log \frac{p(X, Z, c)}{q(c, Z|X)} \right] = \\
\mathbb{E}_q \left[ \sum_{s=1}^S \log p(x_s|z_s, c) \right] + \sum_{l=1}^{L-1} \mathbb{E}_q \left[ \sum_{s=1}^S \log \frac{p(z_l^s|z_{l+1}^s, c, x_s)}{q(z_l^s|z_{l+1}^s, c, x_s)} \right] + \mathbb{E}_q \left[ \sum_{s=1}^S \log \frac{p(z_L|c)}{q(z_L|c, x_s)} \right] - \text{KL}(q(c|X), p(c)) = -\mathcal{L}(X).
\]

**HFSGM.**

\[
\log p(X) \geq \mathbb{E}_{q(c, Z|X)} \left[ \log \frac{p(X, Z, c)}{q(c, Z|X)} \right] = \\
\mathbb{E}_q \left[ \sum_{s=1}^S \log p(x_s|z_s, c) \right] + \sum_{l=1}^{L-1} \mathbb{E}_q \left[ \sum_{s=1}^S \log \frac{p(z_l^s|z_{l+1}^s, c_l)}{q(z_l^s|z_{l+1}^s, c_l, x_s)} \right] + \mathbb{E}_q \left[ \sum_{s=1}^S \log \frac{p(z_L|c_L)}{q(z_L|c_L, x_s)} \right] + \\
\mathbb{E}_q \left[ \sum_{l=1}^{L-1} \log \frac{p(c_l|c_{l+1}, Z_{l+1})}{q(c_l|c_{l+1}, Z_{l+1}, X)} \right] - \text{KL}(q(c_L|X), p(c_L)) = -\mathcal{L}(X).
\]
A.4. Loss
The final loss for all the models is computed per-sample. Given a distribution of $T$ sets (or tasks) of size $S$, the training loss is: $L = \frac{1}{ST} \sum T \mathcal{L}(X_t)$.

A.5. Evaluation
For VAEs we evaluate the models approximating the log marginal likelihood using $S$ importance samples:

$$\text{MLL}(x) = \log \frac{1}{IS} \sum_{i=1}^{IS} \frac{p(x, z)}{q(z|x)} z \sim q(z|x).$$

(21)

For hierarchical models like the NS we use:

$$\text{MLL}(X) = \log \frac{1}{IS} \sum_{i=1}^{IS} \frac{p(X, Z_{is}, c_{is})}{q(Z_{is}, c_{is}|X)} Z_{is} \sim q(Z|c, X), c_{is} \sim q(c|X).$$

(22)

A.6. Learnable Aggregation
In this subsection we describe more explicitly the aggregation mechanism. Given a set $X$, embeddings for samples in the set $h_{s} = f_{\phi}(x_{s})$, and aggregated statistics $r = \frac{1}{S} \sum_{s=1}^{S} h_{s}$, we can compute attention weights for a NS$_{LAG}$ as follows:

$$\alpha(r, h_{s}) = \sigma(\text{dot}(q(r), k(h_{s})))$$

$$r_{LAG} = \sum_{s=1}^{S} \alpha(r, h_{s}) v(h_{s})$$

$$q_{\phi}(c \mid X) = \mathcal{N}(c \mid \mu(r_{LAG}), \sigma(r_{LAG})),$$

(23)

where $q$, $k$ and $v$ are linear layers and $\text{sim}$ is the dot-product scaled by the square root of the representations dimensionality. This approach is inspired by (Lee et al., 2019) and resembles self-attention with a fundamental difference: instead to map from $S$ samples to $S$ samples, we map from $S$ samples to a per-task learnable aggregation. The query input is a handcrafted aggregation (mean, max pooling). We improve the query scoring the handcrafted aggregation with the samples in the set.

For a full SCHA-VAE$_{LAG}$, we can similarly write:

$$r = \frac{1}{S} \sum_{s=1}^{S} f_{\phi}(x_{s}, z_{s}, c)$$

$$\alpha(r, h_{s}, z_{s}, c) = \sigma(\text{dot}(q(r), k(h_{s}, z_{s}, c)))$$

$$r_{LAG} = \sum_{s=1}^{S} \alpha(r, h_{s}, z_{s}, c) v(h_{s}, z_{s}, c)$$

$$q_{\phi}(c_{l} \mid c_{l+1}, Z_{l+1}, X) = \mathcal{N}(c_{l} \mid \mu(r_{LAG}^{l}), \sigma(r_{LAG}^{l})),$$

(24)
A.7. Sampling Algorithms

Algorithm 3 Conditional and Refined Sampling SCHA-VAE

Input: $X$
single pass
$c_L \sim q(c_L|X)$
$z, c_{<L} \sim p(z, c_{<L}|c_L)$
$x \sim p(x|z, c)$
return $x$

Input: $X$
single pass
$c_L \sim q(c_L|X)$
$z, c_{<L} \sim p(z, c_{<L}|c_L)$
$x \sim p(x|z, c)$
repeat
$\tilde{X} = [X, x]$
$\tilde{Z}, c \sim q(\tilde{Z}, c|\tilde{X})$
$\tilde{X}' \sim p(\tilde{X}'|\tilde{Z}, c)$
$x = x'$
until Convergence
return $x$
B. Additional Experiments

B.1. Generalization

Figure 9: Lowerbound in Bits per Dimension (bpd) for CelebA 64x64 test set (unknown classes). We train a CNS, CNS\textsubscript{LAG}, and SCHA-VAE with 3 stochastic layers and process the set information at one single resolution (4x4). We notice that even in this unfavourable scenario, where SCHA-VAE cannot exploit depth and multiresolution aggregation, we gain in terms of likelihood thanks to the hierarchical representation for c.

| CelebA | Depth c | Res c | Depth z | NLL (bpd) |
|-------|---------|-------|---------|-----------|
| CNS   | 1       | 4x4   | 3       | $\leq 4.2235$ |
| CNS\textsubscript{LAG} | 1 | 4x4 | 3 | $\leq 4.2126$ |
| SCHA-VAE | 3 | 4x4 | 3 | $\leq 4.1552$ |

Table 3: Omniglot test negative loglikelihood for convolutional and autoregressive based models. We train a deeper SCHA-VAE with 24/33 stochastic layers and input set size 5. We augment the sets with small translations and rotations. We approximate the likelihood computing the importance weighted bound with S=1000 on homogeneous sets of cardinality 20 (all the samples for a given class/character). The Variational Homoencoder (VHE) uses a PixelCNN to encode c and decode x. Generative Matching Network (GMN) uses RNN layers for encoding/decoding, and attention for aggregation. We show the complexity to compute c, forward $p(X|c)$ and sample $p(x|X)$. We can see that SCHA-VAE is competitive with and without autoregressive components. $n$: set input dimensionality. $d$: image dimension, where in few-shot generation $d >> n$.

| Aggregation | Sampling | NLL (nats) |
|-------------|----------|------------|
| **without autoregressive components** | | |
| VHE (Hewitt et al., 2018) | $O(n)$ | $O(1)$ | 104.67 |
| NS (Edwards & Storkey, 2016) | $O(n)$ | $O(1)$ | 102.84 |
| CNS | $O(n)$ | $O(1)$ | 77.21 |
| SCHA-VAE (Ours) | $O(n)$ | $O(1)$ | **64.34** |
| **with autoregressive components** | | |
| NS + PixelCNN | $O(n)$ | $O(d)$ | 73.50 |
| GMN (Bartunov & Vetrov, 2018) | $O(n^2)$ | $O(n)$ | 62.42 |
| VHE (Hewitt et al., 2018) | $O(n)$ | $O(d)$ | **61.22** |

Figure 10: Set Cardinality. All models are trained with homogeneous (one concept) sets. Lower-bounds for models trained with input set size 5 varying the test set cardinality from 1 to 20 on CelebA. All the models struggle with set size larger than 10 because of the large variety in each set (age, perspective, general look). We can see that learnable aggregation and hierarchical inference both help in modeling the dataset.
B.2. Sampling

Figure 11: Sampling. Different ways to sample the model. (Top): Refined samples obtained using Algorithm 1. Given a small set from an unknown character (right on black background), we sample the model and then refine iteratively using the inference model. We show 20 iterations from left to right. We can see how the generative process refines its guess at each iteration improving c and z in a joint manner. (Middle): Stochastic reconstruction, input sets, conditional sampling, Refined sampling and unconditional sampling (sometimes referred to as imagination) on Omniglot. The models are trained on subsets of Omniglot and tested on disjoint characters. (Bottom): Stochastic reconstruction, input sets, conditional sampling, and unconditional sampling (sometimes referred to as imagination) on CelebA. The models are trained on subsets of CelebA and tested on disjoint identities.
B.3. Transfer

Figure 12: Transfer. Model trained on Omniglot with set size 5 and tested on MNIST, DOUBLE-MNIST and TRIPLE-MNIST (from left to right) with different set size. We can see how our models perform better than a CNS. In particular SCHA-VAE with learnable aggregation (LAG) can adapt better to the new datasets. We test the model transfer capacities on scenarios of increasing complexity, using a subset of disjoint classes from simple out-distribution on MNIST and more challenging out-distribution generalization on variants with 20 and 200 classes.
### B.4. Classification

In this paper our main interest is in improving few-shot generation specifically through the lens of hierarchical inference, multiresolution and learnable aggregation. However is natural to ask how the model perform on downstream tasks, in particular on the supervised few-shot learning task. Here we focus on a simple experiment: we train the models on Omniglot and test on MNIST without any form of adaptation. We consider different ways to approximate a classifier. We can use the fitted generative model as part of a few-shot Bayes classifier:

$$p(y|x, X) = \frac{p(y)p(x|X_y)}{\sum_{y'} p(y')p(x|X_{y'})},$$

where $X = \{X_1, \ldots , X_C\}$ is the set of datasets for the $C$ classes. In Appendix B we investigate two approaches that approximate the predictive distribution $p(x|X_y)$ and one based on $q(c|X)$:

- **ELBO difference:**
  $$\log p(x|X_y) = \log p(x, X_y) - \log p(X_y) \approx \text{ELBO}(x, X_y) - \text{ELBO}(X_y).$$

- **Equation (6):** Sample $q(c|X_y)$ and evaluate:
  $$\mathbb{E}_{q(c)}[\log p(x|c)].$$

- **The classification approach of (Edwards & Storkey, 2016):**
  $$\operatorname{argmax}_y \text{KL}(q(c|X_y), q(c|x)).$$

**Table 4:** Few-shot classification. Models trained on Omniglot and tested on binarized MNIST. Input set dimension 5. For consistency with the KL classifier used in NS, we use only one layer of posterior, the one closer to the data.

|       | AGG | ELBO $[x|X]$ | KL $[q_X, q_x]$ | $\mathbb{E}_q[\log p(x|c)]$ |
|-------|-----|-------------|----------------|--------------------------|
| NS    | MEAN| 0.46        | 0.73           | 0.74                     |
| CNS   | MEAN| 0.41        | 0.76           | 0.76                     |
| CNS   | LAG | 0.33        | 0.75           | 0.76                     |
| SCHA-VAE | MEAN | **0.52** | 0.71 | 0.75 |
|                      | Omniglot          | CelebA           | CIFAR100         |
|----------------------|-------------------|-------------------|------------------|
| **Dimension**        | 1x28x28           | 3x64x64           | 3x32x32          |
| **Number classes**   | 1623              | 6349              | 100              |
| **Classes Train**    | 1000              | 4444              | 60               |
| **Classes Val**      | 200               | 635               | 20               |
| **Classes Test**     | 423               | 1270              | 20               |
| **α**                | $1 \div 2$        | $1 \div 2$        | $1 \div 2$       |
| **α step**           | $0.5 \div 0.98$   | $0.5 \div 0.98$   | $0.5 \div 0.98$  |
| **Batch norm**       | ✗                 | ✗                 | ✗                |
| **Batch size**       | 128               | 16                | 32               |
| **Channels latent c**| 32                | 32                | 32               |
| **Channels latent space** | 128              | 128 $\div 256$  | 128              |
| **Channels latent z**| 32                | 32                | 32               |
| **Classes per set**  | 1                 | 1                 | 1                |
| **Epochs**           | 1000              | 600               | 600              |
| **Heads**            | 4                 | 4                 | –                |
| **Input set size**   | 2–20              | 2–20              | 2–20             |
| **Learning rate**    | $1e^{-3}/2e^{-4}$ | $2e^{-4}$         | $2e^{-4}$        |
| **Schedule**         | plateau/step      | plateau           | plateau          |
| **Likelihood**       | Bernoulli         | DMoL/Bernoulli    | DMoL/Bernoulli   |
| **Optimizer**        | Adam              | Adam              | Adam             |
| **Residual layers**  | 3                 | 3                 | 3                |
| **Resolution latent space** | 1,2,4,8         | 1,2,4,8,16       | 1,2,4,8,16      |
| **Stochastic layers**| 3–12              | 3–24              | 3–24             |
| **Weight decay**     | 0.00001           | 0.00001           | 0.00001          |

### C. Implementation Details

The baseline models are close approximation of the Neural Statistician adapted from: [https://github.com/conormdurkan/neural-statistician](https://github.com/conormdurkan/neural-statistician). The images are encoded using a shared encoder with 3x3 convolutions plus batch normalization. resolution is halved using stride 2. The decoder is the same, with resolution doubled using transposed convolutions. More powerful and expressive decoders can be employed (Oord et al., 2016; Salimans et al., 2017) or multi-resolution deep latent variable models (Maaløe et al., 2019; Vahdat & Kautz, 2020; Child, 2020). We do not use sample dropout (removing random samples from the input set and use the set statistics as additional features).

The loss is a weighted negative lower-bound: $L = -\text{VLB}(\alpha) = (1 + \alpha)\text{REC} + (\text{KL}_z + \text{KL}_c)/(1 + \alpha)$ where alpha is annealed decreasing at each epoch $\alpha = \alpha * \alpha_{\text{step}}$, with $\alpha_{\text{step}} < 1$ at the beginning of training. This re-weighting tends to magnify the importance of the likelihood term and reduce the risk of posterior collapse at the beginning of training. Learning is slower but typically the model and posterior learned are better.

**Modulation.** The way we condition the prior and generative model greatly enhances the adaptation capacities of the model. Other than conditioning the representations through direct concatenation or summation, we can condition directly the features and activations in the residual blocks. FiLM (Perez et al., 2017) is a modulation module used in transfer learning and large-scale class conditional generation. Given an input $c$ and a feature map $f$ with $H$ channels, FiLM learns an affine transformation with $2H$ modulation parameters $(\gamma_h, \beta_h)_{h=1}^{H}$ for each input $c$:

$$\text{FiLM}(z,c) = \gamma(c)f(z) + \beta(c).$$

This affine transformation can be applied in different part of the prior and generative model. In the main paper we use a simplification of such approach, applying $c$ to $z$ using only the bias term, i.e. $\gamma(c) = I$ and $\beta(c) = c$. Empirically we found that for tasks with a unique concept, like sets of characters and faces, this approach is effective.