

A new perspective on probabilistic image modeling

Alexander Gepperth
Applied Computer Science Department
University of Applied Sciences Fulda
Fulda, Germany
alexander.gepperth@cs.hs-fulda.de

Abstract—We present the Deep Convolutional Gaussian Mixture Model (DCGMM), a new probabilistic approach for image modeling capable of density estimation, sampling and tractable inference. DCGMM instances exhibit a CNN-like layered structure, in which the principal building blocks are convolutional Gaussian Mixture (cGMM) layers. A key innovation w.r.t. related models like sum-product networks (SPNs) and probabilistic circuits (PCs) is that each cGMM layer optimizes an independent loss function and therefore has an independent probabilistic interpretation. This modular approach permits intervening transformation layers to harness the full spectrum of (potentially non-invertible) mappings available to CNNs, e.g., max-pooling or half-convolutions. DCGMM sampling and inference are realized by a deep chain of hierarchical priors, where a sample generated by a given cGMM layer defines the parameters of sampling in the next-lower cGMM layer. For sampling through non-invertible transformation layers, we introduce a new gradient-based sharpening technique that exploits redundancy (overlap) in, e.g., half-convolutions. DCGMMs can be trained end-to-end by SGD from random initial conditions, much like CNNs. We show that DCGMMs compare favorably to several recent PC and SPN models in terms of inference, classification and sampling, the latter particularly for challenging datasets such as SVHN. We provide a public TF2 implementation.

I. INTRODUCTION

This conceptual work is in the context of probabilistic image modeling by a hierarchical extension of Gaussian Mixture Models (GMMs), which we term Deep Convolutional Gaussian Mixture Model (DCGMM). Main objectives are density estimation, image generation (sampling) and tractable inference (e.g., image in-painting).

Recent approaches (e.g., GANs, VAEs) excel in image generation by harnessing the full spectrum of CNN transformations, such as convolution or pooling. An issue is however the lack of density estimation and tractable inference capacity, i.e., explicitly expressing and exploiting the learned probability-under-the-model $p(x)$ of an image x . In contrast, recent “deep” probabilistic approaches like sum-product-networks and probabilistic circuits [14], [15], [21], [13], [2] have been explicitly designed to perform these functions on images. However, strong constraints must be satisfied at every hierarchy level to maintain a global probabilistic interpretation, which excludes, e.g., overlapping convolutions and pooling. We aim at overcoming these limitations by constructing deep hierarchies of convolutional Gaussian Mixture Models (GMMs) in order to perform density estimation, realistic sampling and inference within a single model for large-scale, complex visual problems.

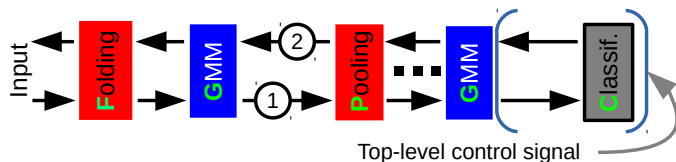


Fig. 1. High-level structure of a typical DCGMM instance with principal layer types. DCGMMs have two operation modes: ①the forward mode which estimates the probability of a presented data sample ②the backwards mode in which a top-level control signal is propagated in reverse direction to generate a sample according to the learned model. The whole architecture is trained end-to-end, with each trainable layer being independently optimized.

A. DCGMM: Model overview and salient points

DCGMM instances can operate in density estimation (“forward”) and sampling (“backwards”) mode, and are realized by a succession of convolutional Gaussian Mixture Model (cGMM) and transformation layers, see Fig. 1.

Most prominently, DCGMMs optimize **independent loss functions** (no back-propagation) for all cGMM layers. This ensures a probabilistic interpretation of cGMM layer outputs and removes the need for structural constraints as imposed in related models (see Sec. I-B).

Since DCGMMs relax the assumption of a single loss function, they can perform **standard CNN operations** between cGMM layers, e.g., half-convolution and pooling, which are non-invertible to different degrees.

In order to allow **sampling through non-invertible operations**, we exploit the fact that DCGMM convolutions are unconstrained and can thus be overlapping as in CNNs. We introduce a novel gradient ascent technique that exploits the redundancy thus created to compensate for the information loss due to non-invertible transformations.

A cGMM layer implements **parameter sharing**, using the same π_k , Σ_k and μ_k for all *positions* of a four-dimensional NHWC input tensor (see Fig. 2). This allows powerful models to be trained with few adjustable parameters.

DCGMM sampling realizes a **deep chain of hierarchical priors**: A cGMM layer propagates posterior probabilities, for which the next upstream cGMM learns a model. A sample drawn from the upstream cGMM thus represents a likely realization of these posteriors. As such, it is a natural choice for controlling sampling in the current layer.

B. Related Work

GANs and VAEs and related models The currently most widely used models of image generation are GANs and VAEs [1], [10], [5], [8]. GANs can sample photo-realistic images [16], but are incapable of density estimation. VAEs show similar performance when it comes to sampling, and outlier detection can be performed, although general density estimation with VAEs is problematic as well. Inference with VAEs is complicated and potentially inefficient. Approaches with similar strength and weaknesses are realized by the PixelCNN architecture [12], GLOW [7] and variants.

Hierarchical MFA/GMM Two related approaches are described in [20], [19]. These approaches realize hierarchical MFA as modular decompositions of single "flat" MFAs, and thus possess a single loss function that is optimized by Expectation-Maximization (EM). Closer to our approach is the proposal of [17], which proposes to train an MFA layer on the inferred latent variables of another, independent MFA instance. Transformations occurring in these models are global, i.e., affect *all* input variables. Local operations like convolutions or max-pooling are not used, and the dimensionality of samples treated in the experiments is low. A preliminary version of DCGMMs was described in [4].

Probabilistic circuits and sum-product networks Probabilistic circuits are deep directed graphs whose leaves compute probabilities from inputs via tractable parameterized distributions. These probabilities are further processed at weighted-sum- and product nodes, with the aim of obtaining a tractable distribution at the root node. This is ensured if some structural constraints are met, and the resulting PCs are often termed Sum-Product Networks (SPNs, see, [15]). In particular, SPNs allow efficient sampling, density estimation and tractable inference even if graphs are very deep.

Similarly to neural networks, finding a graph structure suitable for a particular learning task is challenging, especially if data dimensionality is high. Several interesting ideas for this have been put forward in [14], [15], [21], [13], [2]. An appealing approach is realized by RAT-SPNs [14] which construct random SPNs that are then combined by a single root node, at the price of additional hyper-parameters. Efficient learning

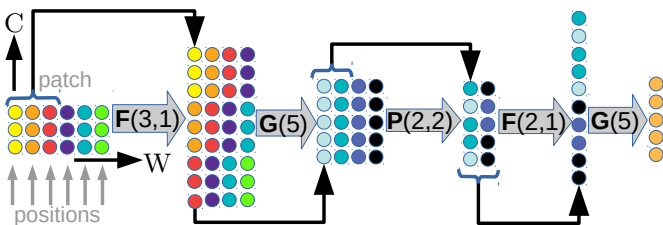


Fig. 2. A simple DCGMM instance applied to a NHWC tensor with N and W dimensions omitted. It contains max-pooling (P), half-convolution/folding (F) and cGMM (G) layers, see text for details. Lower cGMM layers analyze local image *patches* at different *positions*. The topmost cGMM layer is global in the sense that it has a single output position only, in which the whole image is described. Each cGMM layer provides a "fresh" probabilistic interpretation, indicated by uniform coloring for each position.

and inference in SPNs is addressed by Einsum Networks [13] where successive sum and product nodes are combined into a GPU-friendly *Einsum operation*.

SPNs suffer from the structural constraints required for representing valid, tractable distributions. These exclude overlapping convolutions and max-pooling. Convolution can be performed in special cases [2], [21], but general CNN-like convolutions remain inaccessible.

SPNs have been successfully trained on high-dimensional visual problems [14], [21], [14], [2] like the MNIST and FashionMNIST benchmarks. Sampling and inference in SPNs are mostly demonstrated for the "Olivetti faces" benchmark [2], [21], [15] which is high-dimensional but contains only a few hundred samples. To the best of our knowledge, sampling has not been demonstrated for MNIST and FashionMNIST. In [13], sampling and inference is demonstrated for the SVHN benchmark using a simple SPN and extensive data pre-processing. Generated samples are of good quality but not yet comparable to GANs or VAEs. An interesting overview of the research in hierarchical generative mixture models is given in [6].

C. Objective, Contribution and Novelty

The objectives of this article are to introduce a deep GMM architecture which exploits the same principles that led to the performance explosion of CNNs. Novel points are:

- fundamentally new approach to hierarchical mixture modeling based on independent losses
- fundamentally new approach to sampling as a deep chain of hierarchical priors
- use of arbitrary transformations like convolution and pooling in probabilistic image modeling
- realistic sampling for complex visual tasks like SVHN

II. DATASETS

For the evaluation we use the following image datasets:

MNIST [9] consists of 60 000 28×28 gray scale images of handwritten digits (0-9). **FashionMNIST** [22] consists of images of clothes in 10 categories and is structured like the MNIST dataset. **SVHN** [11] contains color images of house numbers (0-9, resolution 32×32).

These datasets are not particularly challenging w.r.t. classification, but their dimensionality of 784 (MNIST, FashionMNIST) and 3072 (SVHN) is high, and the variability of SVHN, in particular, is considerable.

III. METHODS: DCGMM

In order to introduce DCGMMs as a possible hierarchical generalization of GMMs, we will first review facts about vanilla GMMs and introduce the basic notation, which will subsequently be generalized to a multi-layered structure.

A. Review of GMMs

GMMs aim to explain the observed data $X = \{\mathbf{x}_n\}$ by expressing their density as a weighted mixture of K Gaussian component densities $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \equiv \mathcal{N}_k(\mathbf{x}_n)$:

$$p(\mathbf{x}_n) = \sum_{k=1}^K \pi_k \mathcal{N}_k(\mathbf{x}_n), \quad (1)$$

where the normalized component weights π_k modulate the overall influence of each component.

GMMs assume that each observed data sample $\{\mathbf{x}_n\}$ is drawn from one of the K Gaussian component densities \mathcal{N}_k . The selection of this component density is assumed to depend on an unobservable *latent variable* $z_n \in \{1, \dots, K\}$ which follows an unknown distribution. The *complete-data likelihood* for a single data sample reads:

$$p(\mathbf{x}_n, z_n) = \pi_{z_n} \mathcal{N}_{z_n}(\mathbf{x}_n) \quad (2)$$

$$p(\mathbf{X}, \mathbf{z}) = \prod_n p(\mathbf{x}_n, z_n) \quad (3)$$

For a single sample, Bayes' theorem gives us the *GMM posteriors* or *responsibilities* $\gamma(\mathbf{x}_n) = p(z_n = k | \mathbf{x}_n)$:

$$\gamma_k(\mathbf{x}_n) = \frac{p(\mathbf{x}_n, z_n = k)}{\sum_j p(\mathbf{x}_n, z_n = j)} = \frac{\pi_k \mathcal{N}_k(\mathbf{x}_n)}{\sum_j \pi_j \mathcal{N}_j(\mathbf{x}_n)}, \quad (4)$$

which can be computed without the latent variables. Marginalizing the unobservable latent variables out of Eq. (2), we obtain the *incomplete-data log-likelihood* \mathcal{L} :

$$\begin{aligned} \mathcal{L} &= \log p(X) = \log \prod_n p(\mathbf{x}_n) = \log \prod_n \sum_k p(\mathbf{x}_n, z_n = k) \\ &= \log \prod_n \sum_k \pi_k \mathcal{N}_k(\mathbf{x}_n) = \sum_n \log \sum_k \pi_k \mathcal{N}_k(\mathbf{x}_n). \end{aligned} \quad (5)$$

The function \mathcal{L} contains only observable quantities and is a suitable loss function for optimization. At stationary points of Eq. (5), the component weights represent the average responsibilities: $\pi_k = \mathbb{E}_n \gamma_k(\mathbf{x}_n)$. For sampling, one therefore draws a latent variable sample $\hat{z} \sim \mathcal{M}(\boldsymbol{\pi})$ from a multinomial distribution parameterized by the $\boldsymbol{\pi}$, and then draws a sample $\hat{\mathbf{t}} \sim \mathcal{N}_{\hat{z}}$ from the component density $\mathcal{N}_{\hat{z}}$.

B. DCGMM basics

The basic data format in a DCGMM instance (see Fig. 1) are four-dimensional NHWC tensors. For conciseness, we will omit the batch dimension (N) from all formulas. We will denote the dimensions of the current layer L as H, W, C and those of the preceding layer $L-1$ as H', W', C' . Lower-case indices are used similarly, and are sometimes grouped into tuples $\vec{m} = [h, w, c]^T$ or $\vec{m}' = [h', w', c']^T$ for brevity.

Contrary to DNNs, DCGMMs have two operational modes, see Fig. 1: *forward* for density estimation, and *backwards* for sampling. In forward mode, each layer with index $L \geq 1$ receives input from the preceding layer $L-1$, and generates *activities* $\mathbf{A}^{(L)} \in \mathbb{R}^{H,W,C}$ which serve as inputs to the subsequent layer $L+1$. As per the usual convention, $L=0$ denotes the data samples themselves. In backwards mode, each layer L receives

a *control signal* $\mathbf{T}^{(L)} \in \mathbb{R}^{H,W,C}$ from layer $L+1$ and produces another control signal $\mathbf{T}^{(L-1)} \in \mathbb{R}^{H',W',C'}$ for layer $L-1$. We define four layer types, see also App. A: *folding layers* $F(f, \Delta)$ implementing half-convolutions, CNN-like *max-pooling layers* $P(f, \Delta)$ that include a backwards mode, *convolutional GMM layers* $G(K)$ and linear classification layers $C(S)$.

In the following text, we will discuss how the different DCGMM layer types implement the computation of loss functions (where applicable), activities and control signals.

C. Convolutional Gaussian Mixture Layer

cGMM layers are realized by GMMs, slightly modified to a convolutional formulation. As with any GMM (see Sec. III-A), this layer type is defined by K weights $\pi_k^{(L)}$, centroids $\boldsymbol{\mu}_k^{(L)}$ and covariance matrices $\mathbf{P}_k^{(L)}$.

Forward mode Activities $\mathbf{A}^{(L)}$ of the layer are computed as a function of preceding layer activities $\mathbf{A}^{(L-1)}$. Specifically, activities are realized by GMM posteriors, see Sec. III-A. Since the layer is convolutional, the posteriors are computed from the channel content $A_{hw,:}^{(L-1)}$ at every position h, w in the tensor of input activities $\mathbf{A}^{(L-1)}$, see Fig. 2:

$$P_{hw,k}^{(L)}(\mathbf{A}^{(L-1)}) = p(A_{hw,:}^{(L-1)}) \quad (6)$$

$$A_{hw,k}^{(L)}(\mathbf{P}^{(L)}) = \gamma_k(P_{hw,:}^{(L)}) \quad (7)$$

Backwards mode In this mode, the cGMM layer generates a control signal $\mathbf{T}^{(L-1)}$ by sampling at every position h, w , see Sec. III-A. Instead of the parameters $\vec{\pi}$, sampling is governed by an hierarchical prior: the up-stream control signal $\mathbf{T}^{(L)}$. In analogy to Sec. III-A, we can write: $T_{hw,:}^{(L-1)} \sim \mathcal{N}_{\hat{z}_{hw}^{(L-1)}}$, with $\hat{z}_{hw}^{(L-1)} \sim \mathcal{M}(T_{hw,:}^{(L)})$. In case there is no control signal (layer is topmost DCGMM layer), the $\hat{z}_{hw}^{(L-1)}$ are drawn from an uniform distribution.

Loss function For efficient and numerically stable training, we use an approximation to the GMM log-likelihood (see Sec. III-A), which we term the *max-component approximation*, see [3]. It is analogous to EM with hard assignments, see, e.g., [20]. As stated in Sec. I-A, each cGMM layer has a loss function that is computed and optimized independently of other layers. For a single sample, it reads:

$$\mathcal{L}^{(L)} = \mathbb{E}_{h,w} \log \max_k P_{hw,k}^{(L)} \quad (8)$$

An essential point about cGMM layers is that the activities $\mathbf{A}^{(L)}$ and control signals $\mathbf{T}^{(L-1)}$ are computed using a single set of centroids, weights and covariance matrices at every position. Thus, these quantities are **shared** in the same way CNN filters are shared across an image. In this way, large images can be described while requiring relatively few trainable parameters. If memory consumption is not an issue, **independent** parameters can be used as well.

D. Max-Pooling Layer

Pooling is governed by a *kernel size* f and a *stride* Δ , see Fig. 2, where we usually assume non-overlapping pooling: $f=\Delta$. We assign to every position \vec{m} a *receptive field* $\nu(\vec{m})=\{\vec{m}':h'\in[h\Delta,h\Delta+f],w'\in[w\Delta,w\Delta+f],c'=c\}$. In forward mode, this layer type behaves exactly like a CNN max-pooling layer, see Fig. 2: $A_{\vec{m}}^{(L)}=\max_{\vec{m}'\in\nu(\vec{m})}A_{\vec{m}'}^{(L-1)}$. In backwards mode, the pooling layer aims at reconstructing a tensor that would have resulted in the provided control signal. We choose to perform *upsampling* here: $T_{\vec{m}'\in\nu(\vec{m})}^{(L-1)}=T_{\vec{m}}^{(L)}$. The non-uniqueness of this mapping must be counteracted by sharpening, see Sec. III-I.

E. Folding Layer

Forward mode Folding layers perform a half-convolution on preceding layer activities governed by a *kernel size* f and a *stride* Δ , see also Fig. 2. No computation is performed, just a remapping of activities, from position $\vec{m}'(\vec{m})$ in layer $L-1$ to \vec{m} in layer L , see App. A for details on this relation: $A_{\vec{m}}^{(L)}=A_{\vec{m}'(\vec{m})}^{(L-1)}$.

Backwards mode Inverting the one-to-many mapping $\vec{m}'(\vec{m})$ can be done in J different ways $\vec{m}_j(\vec{m}')$. To obtain a control signal, we average over all possibilities: $T_{\vec{m}'}^{(L-1)}=\frac{1}{J}\sum_j T_{\vec{m}_j(\vec{m}')}^{(L)}$.

F. Classification Layer

This layer implements linear classification for S classes. Trainable parameters are the weight matrix \mathbf{W} and the bias vector \mathbf{b} . In forward mode, it generates per-class probabilities as $P_s^{(L)}(\mathbf{x})=S_s(\text{flatten}(A^{(L-1)}\mathbf{W}+\mathbf{b}))$, where S_s represents component s of the softmax function. In backwards mode, it performs an approximate inversion of this operation: $\mathbf{T}^{(L-1)}=\text{reshape}_{H'W'C'}\mathbf{W}^T(\log(\mathbf{T}^{(L)})-\mathbf{b}+c)$. The control signal $\mathbf{T}^{(L)}$ must contain a one-hot encoding of the class that should be generated. The constant c that arises due to the ambiguity in inverting the softmax must be chosen such that the control signal is positive, as it must be if it is to represent GMM posteriors. Classification layers optimize an independent cross-entropy loss function.

G. End-to-end DCGMM training

A defining characteristic of DCGMMs is the fact that each trainable layer optimizes its own loss function. We propose a training scheme where all layers are optimized in parallel. To avoid convergence problems, we start adapting the trainable layer L at time $\delta^{(L)}$, which increases with the position in the hierarchy. Thus, lower layers achieve some convergence before higher layers start their adaptation, which works well in practice, for small delays $\delta^{(L)}$.

Training cGMM layers is performed by SGD from random initial conditions as detailed in [3]. As explained in [3], SGD parameters are very robust and are kept constant throughout all experiments in this article.

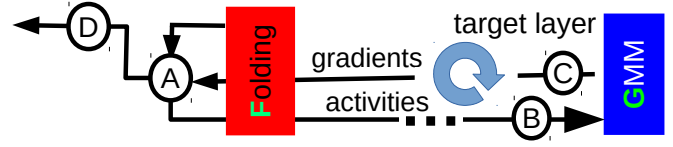


Fig. 3. The gradient-based sharpening procedure, see text for details. Ⓐ Control signal sampled from folding layer Ⓑ forward propagation of control signal to chosen target GMM layer Ⓒ back-propagated gradients applied for modifying control signal Ⓓ final control signal passed to layer $L-1$.

H. Density estimation and outlier detection

These are the principal functions of the forward mode, see Fig. 1. In contrast to, e.g., deep MFA or SPN instances, see Sec. I-B, where sample probability is expressed by the root node, any cGMM layer L expresses sample probability by its log-likelihood $\mathcal{L}^{(L)}$. Lower layers usually model small image patches, whereas higher ones capture the global image structure. We evaluate in Sec. V-B which cGMM layers are best suited for outlier detection.

I. Sampling and sharpening

Sampling is performed in backwards mode, see Fig. 1. Triggered by a top-level control signal, each layer L operates in backwards mode, generating control signals $\mathbf{T}^{(L-1)}$ until the sampling result is read out for $L=0$. Just as density estimation, sampling is efficiently conducted in mini-batches. As detailed in [4], control signals can be thresholded and renormalized to contain only the S highest values, which controls the variability of sampling.

Max-pooling and folding layers perform forward transformations that are not invertible. In backwards mode, this means that they can generate control signals that differ from the true data statistics. To rectify this, we first observe that upstream cGMM layers “know” the statistics of their inputs through training. Additionally, folding layer outputs contain redundancies since their receptive fields usually overlap, which can be exploited.

We therefore propose a statistics-correcting *sharpening* procedure applied at each folding layer L , see Fig. 3. First, we operate the folding layer in backwards mode to obtain the initial control signal $\mathbf{T}^{(L-1)}(i=0)$. We then select a *target layer* $L'>L$ (usually the next-highest cGMM layer) and perform gradient ascent on $\mathbf{T}^{(L-1)}(i)$ for I iterations with step size ϵ_{sh} . Goal is to maximize the target layer loss $\mathcal{L}^{(L')}(i)$ obtained by forward-propagating $\mathbf{T}^{(L-1)}(i)$. In this fashion, we obtain a statistics-corrected control signal $\mathbf{T}^{(L-1)}\equiv\mathbf{T}^{(L-1)}(i=I)$.

J. Tractable Inference: in-painting

In in-painting, we present a partially occluded image and ask the DCGMM instance to complete the missing part. To achieve this, we perform a forward pass up to the topmost cGMM layer X , and then a backwards pass with control signal $\mathbf{T}^{(X)}=\mathbf{A}^{(X)}$. The sampling result for the unknown image part is used to complete the input sample.

IV. METHODS: MODELS AND PARAMETERS

DCGMM We define 7 DCGMM instances which are summarized in App. A. They differ in depth and by their use of stepped folding –vs– max-pooling. The principal hyperparameter is the number of components in cGMM layers, which are chosen such that a batch size of 100 remains feasible. In some experiments, we disable parameter sharing between input positions for specific layers. Learning rates and other SGD-related parameters are kept as described in [4], [3]. Since these parameters do not appear to be task-dependent, we do not vary them in the presented experiments. The training delay (see Sec. III-G) for cGMM layer L is set to $\delta^{(L)}=0.1L$, where $L \in \{1, 2, \dots\}$ indicates the number of cGMM layers below the current one.

RAT-SPN We implemented RAT-SPNs as described in the original publication [14] using the implementation proposed in [14]. As in [13], for every experiment we vary the following parameters: number of distributions per leaf region/number of sums per sum node $I, S \in \{5, 20, 40\}$, split depth $D \in \{1, 5, 9\}$ and number of repetitions $R \in \{10, 25, 40\}$. To speed up training, we employ binomial leaf distributions. Training is conducted for 25 epochs. Variances are constrained as in [14].

Deep generalized convolutional SPNs (DGCSPNs) We implemented the convolutional architecture for generative experiments from the original publication [21] using *libspn-keras*. The number of sums per sum layer is varied as $S \in \{16, 32, 64\}$. Training is conducted for 15 epochs. Accumulators are initialized by a Dirichlet distribution with $\alpha = 0.1$, the number of normal leaf distributions is set to 4, and centroids are initialized from training data as described in [21], where it is also suggested that variances be kept constant at 1.0 for all normal leaves.

Poon-Domingos SPN architecture As an SPN baseline, we use PD-SPN, a very simple instance of the Poon-Domingos architecture as described in [14] using the implementation proposed in [13]. As in [13], we use $\Delta = \{8\}$ for SVHN and $\Delta = \{7\}$ for MNIST, along the horizontal dimension only. We vary the number of distributions per leaf region/number of sums per sum node: $I, S \in \{5, 20, 40\}$.

V. EXPERIMENTS

Compared models are DCGMM, RAT-SPN, PD-SPN and DGCSPN, using parameter (ranges) as given in Sec. IV. All experiments were performed on a cluster of 50 off-the-shelf PCs using nVidia GeForce RTX 2060 GPUs, and our own TensorFlow-based implementation of DCGMMs (see App. A for details). Generally, we repeat each experiment 5 times with different initializations. When grid-searching for feasible parameters, we use the averaged performance measures to identify the best settings. Where feasible, we report mean and standard deviations for experiments.

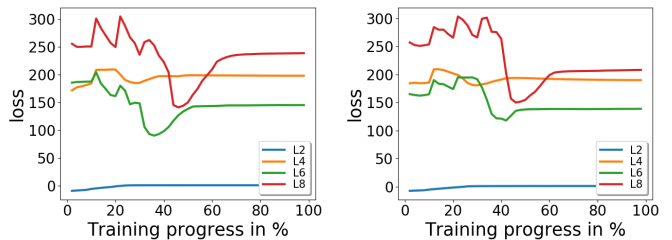


Fig. 4. Test losses for all 4 cGMM layers in the deep DCGMM-F instance for MNIST(left) and FashionMNIST(right). Please note that individual cGMM layers’ losses have different ranges, and that they are maximized (not minimized) by DCGMM training!

ID	MNIST	FashionMNIST	ID	MNIST	FashionMNIST
A	95.4±0.2	62.2±0.3	E	92.8±0.4	74.9±0.2
B	93.3±0.5	72.2±0.2	F	84.1±0.4	68.1±0.1
C	94.2±0.2	74.3±0.1	G	77.4±0.3	62.7±0.1
D	92.5±0.3	68.1±0.1		-	-

TABLE I
OUTLIER DETECTION BY THE TOP LAYERS OF VARIOUS DCGMM INSTANCES, SEE APP. A, MEASURED BY THE MEAN AUC (IN %).

A. DCGMM training dynamics

We train the DCGMM-F instance, see App. A, on the full MNIST and FashionMNIST datasets and record the test losses for each cGMM layer. cGMM Parameters are clamped for a certain percentage of training time in different cGMM layers, see Sec. IV. As we can observe in Fig. 4, cGMM layer converge sequentially in the order of the hierarchy. Initially, losses in layers L4, L6 and L8 decrease due to parameter adaptation in lower layers, before increasing once the latter have become stationary. The fact that final losses are sometimes lower than initial ones underscores that they are not global measures for the whole DCGMM. The individual losses, as depicted in Fig. 4, always converge to the values obtained by a layer-by-layer training scheme.

B. DCGMM outlier detection experiments

The goal of this experiment is to assess the outlier detection capabilities of DCGMMs. In particular, we investigate which DCGMM architectures are most suited for this purpose, and which cGMM layer the outlier detection should be based on. We construct outlier detection problems from classes 1–9 (inliers) –vs– 0 (outliers) of the MNIST and FashionMNIST datasets. Test losses on inlier and outlier classes are recorded for all layers in a DCGMM instance after training on the inliers. Outlier detection is evaluated separately based on each cGMM layer’s loss. By varying the threshold for outlier detection (see Sec. III-H), we obtain ROC-like outlier detection plots, see App. A. The area-under-the-curve (AUC) is used as a quality measure. Results are summarized in Tab. II. We find that the highest cGMM layers are most suited for outlier detection, so we report performance only for these layers. For FashionMNIST, the depth of a DCGMM instance increases its outlier detection capacity. For MNIST, the “flat” DCGMM-A performs best, but MNIST might really be to simple a

ID	MNIST AUC in %	FMNIST AUC in %
DCGMM-E	92.8 ± 0.4	74.9 ± 0.2
RAT-SPN	91.8 ± 0.7	39.2 ± 0.4
DGCSPN	90.6 ± 0.7	57.1 ± 0.9
PD-SPN	91.2± 0.6	48.5 ± 1.7

TABLE II

OUTLIER DETECTION FOR DCGMM-E AND SPN MODELS FOR MNIST AND FASHIONMNIST, QUANTIFIED BY THE AUC MEASURE.

benchmark, and the deep instances have similar performance. Results with different outlier classes are comparable.

C. Outlier detection: model comparison

Here, we compare the outlier detection capacity of DCGMM-E, the best DCGMM instance from Sec. V-B, to various RAT-SPN, PD-SPN and DGCSPN instances, using the same procedures. To find the best parameters for each SPN type, a grid search is conducted over parameters ranges as stated in Sec. IV, repeating each experiment 5 times with identical parameters. Best outlier detection capacities, measured as in Sec. V-B, are reported in Tab. II. We observe a clear edge for DCGMM-E, in particular for FashionMNIST.

D. DCGMM sampling and sharpening

To demonstrate how the sharpening procedure of Sec. III-I improves sampling through non-invertible operations, we train DCGMM instances with max-pooling (C and D, see App. A), to sample from MNIST and FashionMNIST. Then, we compare sampling results with and without sharpening, see Fig. 5. When sharpening is used, target layers are always the next-highest cGMM layers, and gradient ascent is performed for $I = 300$ iterations using a step size $\epsilon_{sh}=1.0$. Figure 5 shows that strong blurring effects occur without sharpening, due to ambiguities in inverting max-pooling layers. In contrast, sharpening removes these ambiguities while maintaining diverse samples. This works best for DCGMM-B (one max-pooling layer), whereas more max-pooling steps seem to destroy too much information, leading to frayed-looking samples. We conclude that "softer" strategies than max-pooling may be required for sampling with really deep DCGMMs. The corresponding figures for FashionMNIST are given in App. A and corroborate this view.

E. Sampling: visual model comparison

In this experiment, we perform a visual comparison of samples generated from DCGMM and the SPN models described in Sec. IV. SPN models are used in the configurations given in the literature (see Sec. IV) and the best parameters found by grid-search in Sec. V-C. Since the configurations in the literature are tailored to certain datasets, we restrict this investigation to the MNIST and FashionMNIST datasets. For DCGMM, we selected DCGMM-F, see App. A, an instance without max-pooling, since this generally leads to more realistic samples, see Sec. V-D.

Typical samples generated by these instances are shown in Fig. 6. We observe that the DCGMM samples generally look smoother and more realistic than SPN-generated ones.

ID	parameters	MNIST acc.	FMNIST acc.
DCGMM-A	38.416	98.2± 0.1	80.3± 2.6
DCGMM-B	293.657	98.7± 0.05	83.6± 1.7
DCGMM-E	40.850	99.1± 0.2	94.5 ± 1.7
DCGMM-F	50.850	99.9 ± 0.07	89.0 ± 2.5
RAT-SPN	1.187.840	99.2± 0.1	85.4 ± 1.9
DGCSPN	7.560.576	98.0± 0.1	80.4± 6.0
PD-SPN	515.683	97.2± 0.2	81.2± 2.4

TABLE III

SAMPLE GENERATION CAPACITY FOR DCGMM AND SPN MODELS AS MEASURED BY A CNN CLASSIFIER ON GENERATED SAMPLES, SEE TEXT FOR DETAILS. WE OBSERVE THAT ALL MODELS PERFORM VERY SIMILARLY ON MNIST, WHEREAS DCGMM-E OUTPERFORMS SPNS FOR FASHIONMNIST. AVERAGES AND STANDARD DEVIATIONS ARE TAKEN OVER 5 INDEPENDENT TRAINING RUNS OF THE CNN CLASSIFIER.

Instances from other DCGMM instances, and the corresponding FashionMNIST samples, are shown in App. A.

F. Sampling: quantitative model comparison

Since visual appearance can be deceptive, this experiment aims to provide a quantitative comparison of the sampling capacity of DCGMMs and SPNs, relying on MNIST and FashionMNIST. We first train a CNN classifier on both datasets to deliver state-of-the-art accuracy for CNNs (details in App. A). DCGMMs and SPNs are separately trained on each class in both datasets, and asked to generate 1000 samples from all 10 classes. The quality measure is the CNN's classification accuracy on the generated samples. Results are shown in Tab. III. We observe that DCGMM sampling seems to produce samples that more accurately match the real data than SPN-based models.

G. Sampling for complex visual problems

Here, we demonstrate that DCGMMs can be trained on complex and high-dimensional visual problems like SVHN, and generate high-quality samples when compared to SPNs. DCGMM instances A and B (the latter with the higher cGMM layer in "independent" mode, see Sec. III-C) are separately trained on each SVHN class, and then used for sampling. We observe that the deeper DCGMM-B instance generates more diverse images, and that DCGMM-B samples even compare favorably to VAE samples (details in App. A).

H. Inference: in-painting

We perform in-painting (see Sec. III-J) on MNIST samples from which the right half was erased. In-painting has the same complexity as sampling, i.e., linear in the number of cGMM layers. Results for the shallow DCGMM-A and the deep DCGMM-F instance are shown in Fig. 8. We observe that completion fits the original image better with a deep DCGMM instance. This is natural since DCGMM-A is essentially a vanilla GMM and can just replicate its limited set of components, with added Gaussian noise. A deep DCGMM instance, in contrast, exhibits much greater variability since each cGMM layer adds (guided) randomness in the choice of the component to sample from.

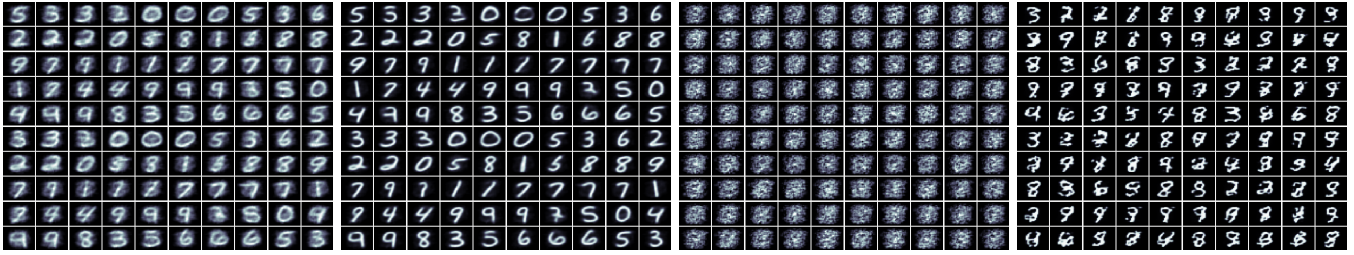


Fig. 5. Samples from DCGMMs with pooling. Left to right: DCGMM-C (no sharpening), DCGMM-C (sharpening), DCGMM-D (no sharpening), DCGMM-D (sharpening). The most beneficial effect of sharpening is observed for the shallow DCGMM-C instance.

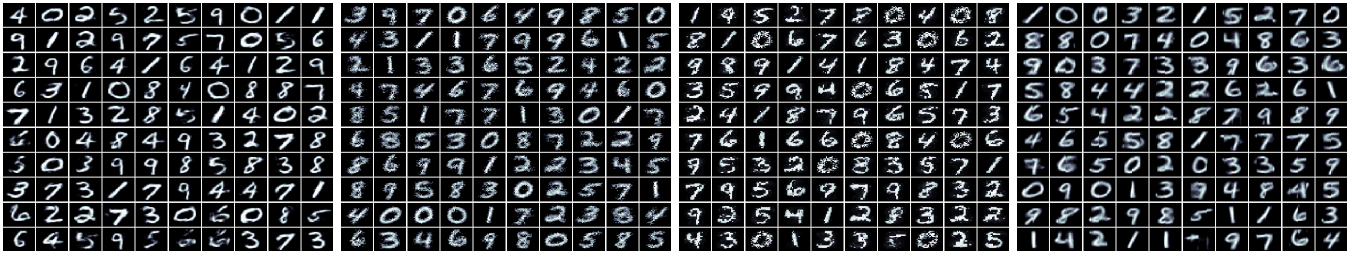


Fig. 6. Visual comparison of sampling results. From left to right: DCGMM-F, RAT-SPN, PD-SPN, DGCSNP.



Fig. 7. SVHN sampling. Left to right: a) SVHN samples b) VAE c) DCGMM-A d) DCGMM-B.

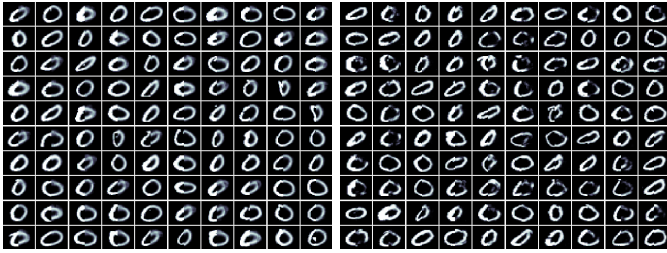


Fig. 8. Examples of in-painting the erased right half of an image. Left: DCGMM-A, right: DCGMM-F.

model	MNIST acc. in %	FMNIST acc. in %
DCGMM-B	98.0 \pm 0.1	89.6 \pm 0.2
RAT-SPN	98.19	89.52
DGCSNP	98.66	90.74

TABLE IV

CLASSIFICATION ACCURACIES IN % OBTAINED FOR MNIST AND FASHIONMNIST. FOR RAT-SPNS AND DGCSNPS, MEAN ACCURACY OVER 5 RUNS IS TAKEN FROM [21].

I. Generative-discriminative learning

This experiment assesses DCGMM classification accuracy on MNIST and FashionMNIST by adding a top-level classifi-

cation layer to the shallow DCGMM-B instance (see Tab. V). To boost classification performance, we use activities of the 2 highest cGMM layers as input to the classifier layer. We find that DCGMM-B classification accuracy is similar but slightly inferior to DGCSNP and RAT-SPN from the literature, see Tab. IV, and that deeper DCGMM instances consistently gave worse results.

VI. PRINCIPAL CONCLUSIONS FROM EXPERIMENTS

Sample probability is best expressed by the topmost cGMM layer loss, see Sec. V-B. This is an important result, since DCGMMs optimize several loss functions, each of which expresses a different local model of the data.

Max-pooling produces excellent results for outlier detection, see Sec. V-B, and is feasible for sampling in shallower DCGMMs due to sharpening, see Sec. V-D). However, in deep DCGMM instances, the information loss seems too severe for good sampling performance. Here, stepped, overlapping half-convolutions are found to be a better choice.

Parameter sharing is very effective for reducing the number of model parameters especially in lower layers, and does not seem to impair sampling performance (see Sec. V-G). This

is probably because the local "visual alphabet" (see App. A) is nearly position-invariant in lower layers and thus can be shared between positions with little loss.

Classification Since DCGMMs optimize independent loss functions, discriminative training of a classification layer does not influence unsupervised training in lower layers, because no back-propagation is performed. Thus, the DCGMM is still be able to generate realistic samples, at the price of similar but slightly inferior classification accuracy, see Sec. V-I. This is in contrast to discriminative SPN training, see [21], [14]. Here, RAT-SPNs and DGC-SPNs perform classification by optimizing a global cross-entropy loss, but can no longer generate realistic samples.

Realistic sampling The DCGMM samples presented in Secs. V-D and V-G have a clear edge w.r.t. samples produced by SPNs. This shows that linking independent probabilistic descriptions (the cGMM layers) by a deep chain of hierarchical priors is a feasible way to describe complex (image) distributions. SPN training on complex problems like SVHN is described in [13], where the PD-SPN architecture (see Sec. IV) is trained on preprocessed SVHN data. Due to non-overlapping scopes in the PD architecture, all samples exhibit "stripe" artifacts which are absent from DCGMM samples, presumably because their scopes *can* overlap.

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[2] C. J. Butz, J. S. Oliveira, A. E. dos Santos, and A. L. Teixeira. Deep convolutional sum-product networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3248–3255, 2019. 1, 2

[3] A. Gepperth and B. Pfülb. Gradient-based training of gaussian mixture models for high-dimensional streaming data. *Neural Processing Letters*, 2021. 3, 4, 5

[4] A. Gepperth and B. Pfülb. Image modeling with deep convolutional gaussian mixture models. In *International Joint Conference on Neural Networks(IJCNN)*, 2021. 2, 4, 5

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 3(January):2672–2680, 2014. 2

[6] P. Jaini, P. Poupart, and Y. Yu. Deep homogeneous mixture models: Representation, separation, and approximation. In *NeurIPS*, pages 7136–7145, 2018. 2

[7] D. P. Kingma and P. Dhariwal. Glow: generative flow with invertible 1×1 convolutions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10236–10245, 2018. 2

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2014. 2

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. 2

[10] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. pages 1–7, 2014. 2

[11] Y. Netzer and T. Wang. Reading digits in natural images with unsupervised feature learning, 2011. 2

[12] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016. 2

[13] R. Peharz, S. Lang, A. Vergari, K. Stelzner, A. Molina, M. Trapp, G. Van den Broeck, K. Kersting, and Z. Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *International Conference on Machine Learning*, pages 7563–7574. PMLR, 2020. 1, 2, 5, 8

[14] R. Peharz, A. Vergari, K. Stelzner, A. Molina, X. Shao, M. Trapp, K. Kersting, and Z. Ghahramani. Random sum-product networks: A simple and effective approach to probabilistic deep learning. In *Uncertainty in Artificial Intelligence*, pages 334–344. PMLR, 2020. 1, 2, 5, 8

[15] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011. 1, 2

[16] E. Richardson and Y. Weiss. On GANs and GMMs. *Advances in Neural Information Processing Systems*, 2018-December(NeurIPS):5847–5858, 2018. 2

[17] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep mixtures of factor analysers. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1:505–512, 2012. 2

[18] M. S. Tanveer, M. U. K. Khan, and C.-M. Kyung. Fine-tuning darts for image classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4789–4796. IEEE, 2021. 9

[19] A. Van Den Oord and B. Schrauwen. Factoring variations in natural images with deep Gaussian mixture models. *Advances in Neural Information Processing Systems*, 4(January):3518–3526, 2014. 2

[20] C. Viroli and G. J. McLachlan. Deep Gaussian mixture models. *Statistics and Computing*, 29(1):43–51, 2019. 2, 3

[21] J. Wolfshaar and A. Pronobis. Deep generalized convolutional sum-product networks. In *International Conference on Probabilistic Graphical Models*, pages 533–544. PMLR, 2020. 1, 2, 5, 7, 8

[22] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. pages 1–6, 2017. 2

APPENDIX

Precise layer configurations of the various DCGMM instances used in this article are given in Tab. V. As a rule, cGMM components were chosen as high as possible while respecting memory constraints during training. All cGMM layers are assumed to use parameter sharing, otherwise the number of trainable parameters will be higher. We observe that the number of trainable parameters actually decreases as more cGMM layers are added. As with CNNs, this is because the number of parameters mainly scales with the filter size of folding layers, which can be kept small in deep architectures. DCGMM layers transform inputs of dimension H', W', C'

ID	Configuration	parameters
A	F(28,1)-(49)	38416
B	F(8,2)-G(49)-F(11,1)-G(49)	293657
C	F(8,1)-G(49)-P(2,2)-G(49)	243236
D	F(3,1)-G(25)-P(2,2)-F(4,1)-G(25)-P(2,2)-F(5,5)-G(49)	40850
E	F(3,1)-G(25)-F(4,2)-G(25)-F(12,1)-G(49)	186625
F	F(3,1)-G(25)-F(4,2)-G(25)-F(4,2)-G(25)-F(5,1)-G(49)	50850
G	F(3,1)-G(25)-P(2,2)-F(3,1)-G(25)-P(2,2)-F(3,1)-G(25)-P(2,2)-F(2,1)-G(49)	16375

TABLE V

OVERVIEW OF DCGMM CONFIGURATIONS USED IN THE EXPERIMENTS. LAYER TYPES ARE F (HALF-CONVOLUTION LAYER), G (CGMM LAYER) AND P(MAX-POOLING LAYER). OPTIONALLY, A LINEAR CLASSIFIER LAYER CAN BE ADDED AT THE TOP OF EACH INSTANCE FOR CONDITIONAL SAMPLING.

into activities of dimension H, W, C . Since each layer in

Layer type	Notation	H	W	C
Folding	$F(f, \Delta)$	$1 + \frac{H'-f}{\Delta}$	$1 + \frac{W'-f}{\Delta}$	$f^2 C'$
Max-Pooling	$P(f, \Delta)$	$1 + \frac{H'-f}{\Delta}$	$1 + \frac{W'-f}{\Delta}$	C'
Classification	$C(S)$	1	1	S
cGMM	$G(K)$	H'	W'	K

TABLE VI

OVERVIEW OF DCGMM LAYER TYPES AND THEIR NOTATION. THE THREE RIGHTMOST COLUMNS INDICATE THE SHAPE OF ACTIVITIES IN FORWARD MODE IF THE LAYER RECEIVES AN INPUT OF DIMENSIONS H', W', C' . FOLDING AND MAX-POOLING LAYERS ARE PARAMETERIZED BY KERNEL SIZE f AND THE STRIDE Δ , GMM LAYERS BY THE NUMBER OF COMPONENTS K AND CLASSIFICATION LAYERS BY THE NUMBER OF CLASSES S .

forward mode implements a deterministic transformation, the dimensions of activities depend only on the dimensions of the inputs as listed in Tab. VI.

The CNN classifier used in Sec. V-F has a layer structure as given in Tab. VII. It is implemented in TensorFlow2/Keras and trained for 15 Epochs on either MNIST or FashionMNIST, using an Adam optimizer, a learning rate of 0.01 and a batch size of 100. For MNIST, this is sufficient for state-of-the-art performance ($> 99\%$), whereas for FashionMNIST, a performance of roughly 91% can be reached. While this is inferior to the performance obtained by more refined models (e.g., [18]), we accept it here for simplicity, and also because the CNN classifier is just a tool to detect differences in the distributions of real and generated samples.

Type	Kernel	Prob.	channels/neurons
Dropout		0.1	
Conv/ReLU	3		64
Conv/ReLU	3		64
Pooling	2		
Conv/ReLU	3		64
Pooling	2		
Conv/ReLU	3		64
Pooling	2		
Dense/ReLU			350
Dropout		0.1	
Dense/ReLU			350
Dense/ReLU			350
Dense/Softmax			10

TABLE VII

HYPER-PARAMETERS OF THE CNN USED FOR ASSESSING SAMPLING PERFORMANCE.

Sharpening behaves similarly for the FashionMNIST dataset as it was found for MNIST in Sec. V-D. Namely, the shallower DCGMM-B instance profits strongly from a sharpening through a single max-pooling layer, but we observe deterioration of sampling performance when more max-pooling layers are involved, as in instance DCGMM-D. Figure 9 shows this quite nicely.

Outlier detection is quantified using a ROC-like curve, plotting kept inliers against rejected outliers while varying the separation threshold that is applied to the log-likelihoods. Typical examples of such curves are shown in Fig. 10.

Table VIII gives details about the convolutional VAE used to generate SVHN samples. It was trained for 100 epochs on all SVHN classes using the Adam optimizer and a learning rate

of 0.0001.

Type	Kernel	Stride	channels/size
Encoder			
Conv/ReLU	3	1	64
Conv/ReLU	3	2	128
Conv/ReLU	3	2	256
Conv/ReLU	2	2	512
Flatten	-	-	4096
Dense			2*64
Decoder			
Dense/ReLU			4*4*128
Reshape			4,4,128
Conv2D ^T /ReLU	4	1	1024
Conv2D ^T /ReLU	5	1	512
Conv2D ^T /ReLU	4	2	256
Conv2D ^T /ReLU	5	1	128
Conv2D ^T /ReLU	3	1	128
Conv2D ^T /ReLU	3	1	3

TABLE VIII

HYPER-PARAMETERS OF THE VAE USED FOR SVHN SAMPLING.

This appendix gives MNIST sampling results in a more complete fashion, that is, including more DCGMM instances, in Fig. 12. Figure 13 gives samples from the same DCGMM instances for FashionMNIST, whereas Fig. 11 shows FashionMNIST samples generated by SPNs.

The lower cGMM layers of a cGMM instance usually model small image patches extracted by a preceding folding layer. With parameter sharing enabled, the cGMM therefore describes all positions within an image using a single set of parameters. This makes the most sense for low-hierarchy layers, since local image content tends to be similar across image at small patch sizes. The cGMM prototypes there form a kind of "visual alphabet", a set of centroids that, together, best describe local image content. We exemplarily show this for MNIST and FashionMNIST by visualizing the centroids of the lowest cGMM layer in instance DCGMM-B, which models 8x8 image patches. We observe in Fig. 14 that the basic building blocks of both datasets are faithfully represented. For implementing RAT-SPN and PD-SPN, we made use of the public code provided under <https://github.com/cambridge-mlg/EinsumNetworks> which relies mainly on PyTorch. DGCSPNs are implemented using *libspn-keras* which is TensorFlow2-based and can be obtained from <https://github.com/pronobis/libspn-keras>. VAEs and CNNs are self-implemented in TensorFlow2/Keras. TensorFlow2-Code for DCGMM can be found under <https://github.com/anon-scientist/ijcnn22-dcgmm>. Code for selected experiment of this article is available under <https://github.com/anon-scientist/ijcnn22-experiments>.

The relation between preceding and current layer activities is governed by a one-to-many mapping. This means that a single activity in layer $L-1$ can be mapped to several activities in layer L by the relation $\vec{m}'(\vec{m})$. One-to-many situations always occur when filters are set to overlap. The precise form

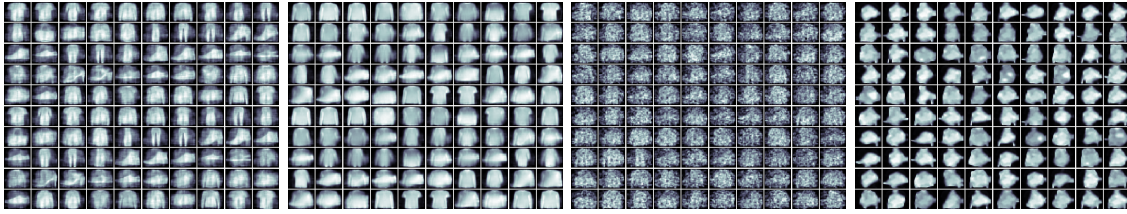


Fig. 9. Sharpening for DCGMMs with pooling on FashionMNIST. Left to right: DCGMM-C (no sharpening), DCGMM-C (sharpening), DCGMM-D (no sharpening), DCGMM-D (sharpening). The most beneficial effect of sharpening is observed for the shallow DCGMM-C instance.

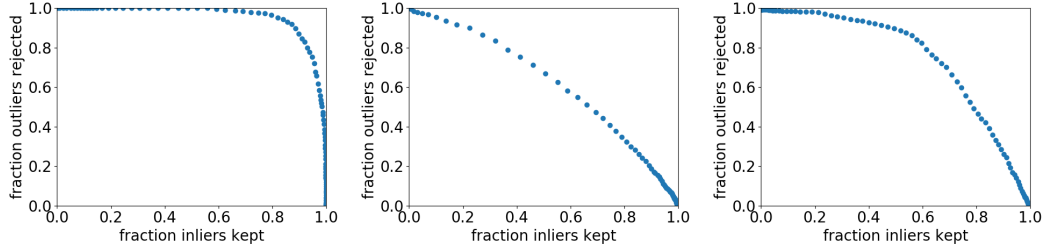


Fig. 10. Examples of ROC-like curves for outlier detection. Left: DCGMM-A (MNIST), middle: DCGMM-A (FashionMNIST), right: DCGMM-E (FashionMNIST). The area under these curves is taken to be a measure of outlier detection capacity.

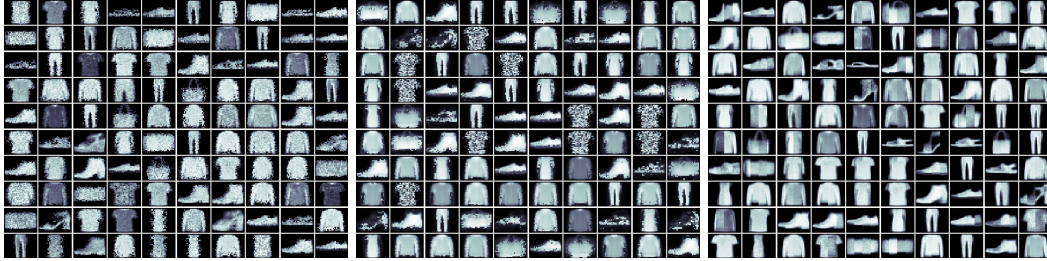


Fig. 11. SPN samples for FashionMNIST. From left to right: RAT-SPN, PD-SPN, DGCSPN.

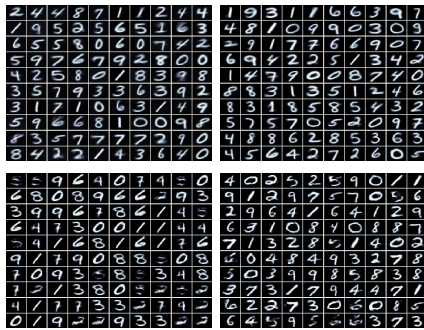


Fig. 12. Samples from several DCGMM instances for MNIST. Upper row: DCGMM-A(left), DCGMM-B(right). Lower row: DCGMM-E(left), DCGMM-F(right).

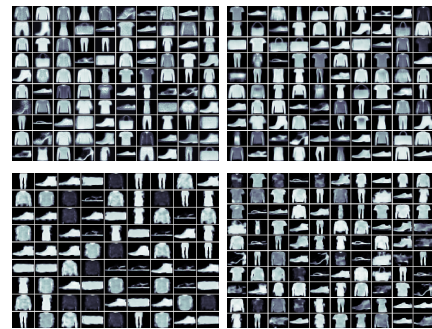


Fig. 13. Samples from several DCGMM instances for FashionMNIST. Upper row: DCGMM-A (left), DCGMM-B (right). Lower row: DCGMM-E (left), DCGMM-F (right).

of this relation reads:

$$\vec{m}'(\vec{m}) = \begin{pmatrix} h\Delta + c // (fC') \\ w\Delta + (c // C') \% f \\ c \% C' \end{pmatrix}, \quad (9)$$

$$(10)$$

where $//$ and $\%$ represent integer and modulo division.

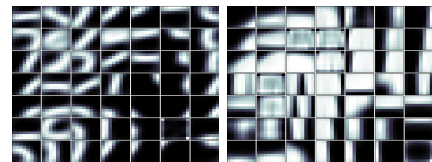


Fig. 14. Visualization of centroids in the lowest cGMM layer of DCGMM-B. Left: training on MNIST, right: training on FashionMNIST.