Deep Part-Based Generative Shape Model with Latent Variables

Alexander Kirillov[1]
alexander.kirillov@tu-dresden

Mikhail Gavrikov[2]
gavrmike@gmail.com

Ekaterina Lobacheva[4]
elobacheva@hse.ru

Anton Osokin[3]
anton.osokin@inria.fr

Dmitry Vetrov[4]
vetrovd@yandex.ru

[1] TU Dresden,
Dresden, Germany

[2] Rubbles,
Moscow, Russia

[3] INRIA – École Normale Supérieure,
Paris, France

[4] National Research University
Higher School of Economics (HSE),
Moscow, Russia

Models of shape play substantial role in a number of computer vision tasks such as segmentation, inpainting and detection. The Shape Boltzmann Machine (SBM) [2] and its multilabel version MSBM [1] have been recently introduced as deep generative models that capture the variations of an object shape. While being more flexible MSBM requires datasets with labeled parts of the objects for training (Fig. 1c). In the paper we present an algorithm for training MSBM using binary masks of objects (Fig. 1b) and the seeds which approximately correspond to the locations of objects parts (Fig. 1d). The latter can be obtained from part-based detectors in an unsupervised manner. We derive a latent variable model and an EM-like training procedure for adjusting the weights of MSBM using a deep learning framework. We show that the model trained by our method outperforms SBM in the tasks related to binary shapes and is very close to the original MSBM in terms of quality of multilabel shapes.

SBM and MSBM SBM is a Deep Boltzmann Machine (DBM) [4] with special constraints on its parameters [2] that allow to avoid overfitting for small datasets. It defines a joint distribution $p(\boldsymbol{b}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta)$, where $\boldsymbol{b}$ is a layer that corresponds to the binary shape of an object (fig. 1b), $\boldsymbol{h}^1, \boldsymbol{h}^2$ are two hidden layers and $\boldsymbol{\theta}$ is a vector of all the SBM parameters. MSBM is a generalization of SBM model to multilabel case and it defines the similar distribution $p(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta)$, where $\boldsymbol{m}$ is a layer which corresponds to the multilabel shape of an object (fig. 1c). MSBM is more expressive since the variations of an object's parts are usually smaller than the variation of the whole object. Both SBMs and MSBMs are trained in exactly the same way as general DBMs [4] using variational EM-algorithm. An important constraint of the scheme is that it requires the full annotation. This is a serious constraint, because obtaining the ground-truth segmentation (especially for multilabel tasks) is not an easy task.

Our model In this paper we make a step towards the unsupervised training of a shape model. Specifically, we propose a way to train a multilabel model without using the full multilabel annotation (as in [1]). Instead we use easier-to-obtain binary masks and seeds $\boldsymbol{s}$ of the parts (fig. 1d). For each part there is one seed pixel that approximatelly corresponds to the center of this part.

We model the joint distribution $p(\boldsymbol{b}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta)$ of binary mask $\boldsymbol{b}$, seeds $\boldsymbol{s}$, multilabel masks $\boldsymbol{m}$ and hidden variables $\boldsymbol{h}^1$, $\boldsymbol{h}^2$ using the assumption that binary segmentation $\boldsymbol{b}$ and seeds $\boldsymbol{s}$ are conditionally independent given multilabel segmentation $\boldsymbol{m}$ (Fig. 2), i.e.

$$p(\boldsymbol{b}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta) = p(\boldsymbol{b}|\boldsymbol{m})p(\boldsymbol{s}|\boldsymbol{m})p(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta). \quad (1)$$

Distribution $p(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2|\theta)$ is represented by an MSBM. For conditional distribution $p(\boldsymbol{b}|\boldsymbol{m})$ we use an intuitive assumption that if a pixel belongs to any part of an object, then it belongs to the object, otherwise it belongs to the background. W.r.t. $p(\boldsymbol{s}|\boldsymbol{m})$ we assume that each pixel impacts the seed positions independently, and that pixel i belonging to the part of object m_i only attracts seed s_{m_i} of this part to be close to the position $f_{coord}(i)$ of the pixel i: $p(s|m) \propto \prod_{i:m_i \neq 0} \mathcal{N}\left(s_{m_i}|f_{coord}(i), \sigma^2\right)$, where σ is the external parameter of the method.

To train the unknown parameters $\boldsymbol{\theta}$ of the model (1) we use the variational EM algorithm, i.e. maximize $\sum_{d=1}^{D} \log p(\boldsymbol{b}^d, \boldsymbol{s}^d|\boldsymbol{\theta})$ w.r.t. parameters $\boldsymbol{\theta}$, where $\{\boldsymbol{b}^d\}_{d=1}^D$ are binary shapes in a training set and $\{\boldsymbol{s}^d\}_{d=1}^D$ are corresponding seeds. At the E-step we obtain distributions $q^d(\boldsymbol{m}, \boldsymbol{h}^1, \boldsymbol{h}^2)$ in the fully-factorized family.

The seeds can be either set manually or found automatically using part-based detectors [3]. Together with the part-based detector (only bounding box annotations required for training) our framework provides the ability to exploit the benefits of multi-part shape models having only the dataset of images with binary object masks annotated.

(a) (b) (c) (d)

Figure 1: (a) – image of an object, (b) – binary segmentation $\boldsymbol{b}$, (c) – the multilabel segmentation $\boldsymbol{m}$, (d) seeds $\boldsymbol{s}$ for the selected 4 parts: head, front legs, rear legs and croup.

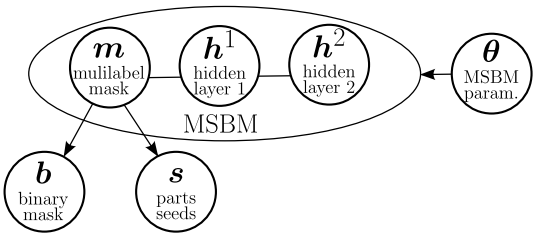

Figure 2: The graphical illustration of our joint model.

Experiments We use several baselines for an experimental evaluation of our approach: the SBM model, the MSBM model trained given the full multilabel annotations by the original method of [1] and MSBM models thained by the same method using the multilabel annotations created automatically from the binary annotations and the seeds (here we associate each pixel with the part whose seed is closest to the pixel in the image plane).

We perform three main experiments. Firstly, we qualitatively compare shape samples generated by MSBMs trained with different procedures. Secondly, we conduct the "imputation score" experiment: we divide each test image into several segments and estimate the conditional probability of the binary ground-truth mask of one segment given the mask of the remaining ones. Finally, we experiment with a task of generating binary shapes if only seeds are known. All the experiments are performed on two datasets: the Weizmann horses and the Caltech-101 motorbikes.

The experiments show that as the model of binary shape the MSBM trained by our technique performs similar to the MSBM that is trained using full annotation and significantly outperforms the SBM and MSBM that is trained using automatically obtained multilabel annotation from a binary mask and seeds. Unlike MSBM trained with full annotation our MSBM generates rather noisy inner boundaries. It occurs since our model has not seen any boundaries between parts at all during the training procedure. From the evaluation perspective this effect is not important if we are interested in a binary object mask.

[1] S. M. Ali Eslami and Chris Williams. A generative model for parts-based object segmentation. In *NIPS*. 2012.

[2] S. M. Ali Eslami, Nicolas Heess, Christopher K. I. Williams, and John Winn. The shape boltzmann machine: a strong model of object shape. In *IJCV*, 2013.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32(9): 1627–1645, 2010.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.