Eurographics Workshop on 3D Object Retrieval (2019), pp. 1–2
S. Biasotti and G. Lavoué (Editors)

Conditional Variational Autoencoders for Image Based Scene Retrieval

L. A. Pérez Rey[1], M. Holenderski[1] and D. Jarnikov[1]

[1] Eindhoven University of Technology

1. Overview

The proposed approach consists of image to image comparison with conditional variational autoencoders (CVAE) [KRMW14]. The CVAE is a semi-supervised method for approximating the underlying generative model that produced a set of images and their corresponding class labels in terms of the so-called unobserved latent variables. Each of the input images is described in terms of a probability distribution over the latent variables and the classes.

Our approach consists of using the probability distributions calculated by the CVAE for each image as a descriptor. The comparison between an image query and the 3D scene renders is with respect to the probability distributions obtained from the CVAE. The method consists of data pre-processing, training and retrieval described in the following subsections.

2. Data Preprocessing

Thirteen renders are obtained for each of the 3D scenes. Each of the 3D scenes has a predefined view when loaded into the SketchUp software. This view is saved as a 2D render together with twelve views at different angles around the scene as in [SMKLM15].

The training data set consists of the 3D scene renders together with the training images. All images are resized to a resolution of 64×64 and all pixel values are normalized to the interval $[0,1]$. Image augmentation is carried out by performing a horizontal flip to all images. The corresponding data space is $X = [0,1]^{64\times64\times3}$.

3. Training

The CVAE consists of an encoder and a decoder neural network. The encoder network calculates from an image $x \in X$ the parameters of a probability distribution over the latent space $Z = \mathbb{R}^d$ and over the thirty class values in $Y = \{1,2,3,\ldots,30\}$. The decoder network calculates from a latent variable $z \in Z$ and a class $y \in Y$, the parameters of a distribution over the data space X .

The distributions for the encoder correspond to a normal distribution over Z and a categorical distribution over Y. A normal distribution over X is chosen for the decoder. The probabilistic model used corresponds to the M2 model described in the article [KRMW14]. Both, the encoding and decoding neural networks are convolutional.

The CVAE is fed with batches of labelled images during training. The loss function is the sum of the negative Evidence Lower Bound (ELBO) and a classification loss. The ELBO is approximated by means of the parametrization trick described in [KRMW14,KW14] and represents the variational inference objective. The classification loss for our encoding distributions over Y corresponds to the cross entropy between the probability distribution over Y with respect to the input label.

4. Retrieval

After training, an image $x \in X$ can be described as a conditional joint distribution over $Z \times Y$. The density $q_\phi(z|x)$ corresponds to a normal distribution and $q_\phi(y|x)$ to a categorical distribution over Y where ϕ represents the weights of the encoder neural network. The joint density corresponds to $q_\phi(z,y|x) = q_\phi(z|x)q_\phi(y|x)$.

The similarity D between an input query image $x^* \in X$ and a 3D scene in terms of its N rendered images $S = \{x_r\}_{r=1}^N$ is given by the minimum symmetrized cross entropy H_s between the query and the render probability distributions, see Figure **??**.

$$D(x^*,S) \min_{r\in\{1,2,\ldots,13\}} H_s(q_\phi(z|x^*), q_\phi(z|x_r)) + \alpha H_s(q_\phi(y|x^*), q_\phi(y|x_r)). \quad (1)$$

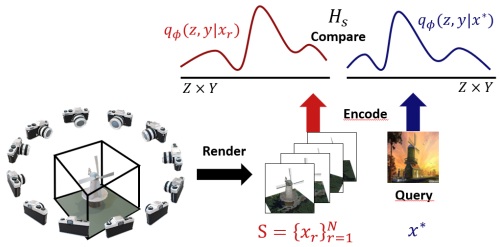
We have used the parameter $\alpha = 64 \times 64 \times 3$ to increase the importance of label matching. A ranking of 3D scenes is obtained for each query according to this similarity.

5. Submissions

We have sent five submissions corresponding to methods who differ only on the architecture of the encoding and decoding neural networks. These are described as follows:

1. **CVAE-(1,2,3,4)**: CVAE with different CNN architectures for the encoder and decoder.

2. **CVAE-VGG**: CVAE with features from pre-trained VGG [Kal17] on the Places data set [ZLK*18] as part of the encoder.



For all figures please keep in mind that you **must not** use images with transparent background!

Figure 1: *Here is a sample figure.*

References

[Kal17] KALLIATAKIS G.: Keras-vgg16-places365. https://github.com/GKalliatakis/Keras-VGG16-places365, 2017. 2

[KRMW14] KINGMA D. P., REZENDE D. J., MOHAMED S., WELLING M.: Semi-Supervised Learning with Deep Generative Models. 1–9. 1

[KW14] KINGMA D. P., WELLING M.: Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)* (2014). doi:10.1051/0004-6361/201527329. 1

[SMKLM15] SU H., MAJI S., KALOGERAKIS E., LEARNED-MILLER E.: Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)* (12 2015), vol. 2015 Inter, IEEE, pp. 945–953. doi:10.1109/ICCV.2015.114. 1

[ZLK*18] ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A., TORRALBA A.: Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 6 (2018), 1452–1464. doi:10.1109/TPAMI.2017.2723009. 2