Unsupervised Learning of Disentangled and Interpretable Representations of Material Appearance

Santiago Jiménez, Julia Guerrero-Viu, Belén Masiá

Graphics & Imaging Lab (GILab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: s.jimenez@unizar.es

Abstract

As humans, we have learned through experience how to interpret the visual appearance of materials in our environment, enabling us to predict the properties of an object just by its looks. Analogously, we propose a learning-based algorithm capable of effectively disentangling certain perceptual features of images in an unsupervised way.

Introduction

Visual perception of materials is key to effectively model and interact with our surroundings. Consequently, finding underlying representations of material appearance that are intuitive for humans can significantly improve applications like material search, retrieval, classification, or editing, enabling a more natural interaction with tools for visual content creation. In this work, we exploit unsupervised learning techniques in order to find latent representations of materials in image space that are both disentangled and interpretable with respect to perceptual factors. By letting our model learn the underlying statistical structure in images without any prior knowledge, we can broaden the spectrum of factors that it can find, in contrast to previous work that analyzes specific material properties, such as glossiness [4]. Also, as opposed to previous supervised approaches [3, 5], we avoid biasing the model to learn the factors that are already annotated by humans, and mitigate the need for gathering large amounts of human annotations.

Method

Reducing data dimensionality is a long-standing problem in computer science. A typical solution to it has been manually programming algorithms that reduce this dimensionality, but using deep learning methods, we can achieve a codification that takes into account abstract features of data. To achieve this, we chose to use the architecture of a FactorVAE [2]. It is a modification of the Variational AutoEncoder (VAE) [1], a network that is composed of an encoder E which translates the input data $x \in R^N$ to a compressed latent representation $z \in R^K | K \ll N$, and a decoder D, that generates a reconstruction of the data $\hat{x} \in R^N$ given its representation in this latent space. During the training, the network learns how to, in an unsupervised way, store relevant information of the data (in our case, perceptual features) in the bottleneck that we call latent space. Furthermore, a FactorVAE network uses an additional component whose task is to maintain a good regularization in the latent space. Figure 1 contains a visual representation of the network and a simplified version of the objective function used to train it.

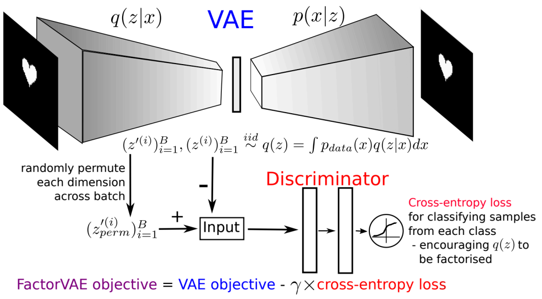
$q(z|x)$
VAE
$p(x|z)$
$(z'^{(i)})_{i=1}^B, (z^{(i)})_{i=1}^B \overset{iid}{\sim} q(z) = \int p_{data}(x) q(z|x) dx$
randomly permute each dimension across batch
-
$(z'^{(i)}_{perm})_{i=1}^B$
+
Input
Discriminator
Cross-entropy loss for classifying samples from each class - encouraging $q(z)$ to be factorised
FactorVAE objective = VAE objective - $\gamma \times$ cross-entropy loss

Figure 1: Architecture of the FactorVAE [2] used in this project. Note the discriminator component, in charge of keeping a well-distributed latent space.

Data Used

The main dataset used during this project is the Serrano Dataset [3]. This set was built originally to study the effects of illumination and geometry in the material perception of objects comprising complex appearances. Adapting the dataset to our needs, we came up with a set of RGB images with 256x256 resolution, containing the representation of a combination of two geometries, nine illuminations, and 520 materials. We also remove the background information, leaving a plain black color, facilitating the convergence of the learning model. Figure 2 contains a sample set of the data used in this work.

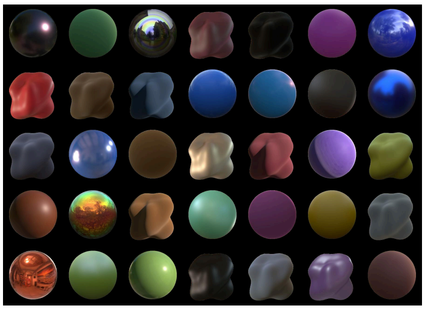

Figure 2: Representative subset of the Serrano Dataset [3] used in this project. We can observe how the samples feature different levels of complexity, both in terms of illumination and material.

Results

Although with the encoder E we managed to reduce drastically the dimensionality of the data, the current dimensionality remains too large to be visualized in a plot. To deal with this, we use the *t-Distributed Stochastic Neighbor Embedding* algorithm (TSNE) to project the latent representations to a 2D space. Then, we color each sample according to a certain label (like *illumination* or *geometry*), which allows us to see how the model clusters data attending to these factors. With this approach, we obtain the plots available in Figure 3, where we can visually assess how the model is effectively creating clusters of data, attending to their illumination and geometry.

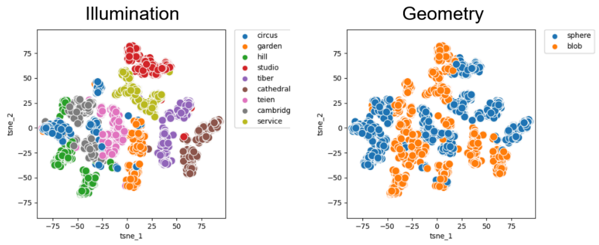

Figure 3: 2D visualization of the latent space of our model, by using TSNE. Every point represents the latent representation of an image, colored according to their illumination (left) and geometry (right).

In order to assess which information is being stored in each of the dimensions in the latent space, we use the *prior traversals* plot. This visualization is obtained by manually creating certain latent representations z, and introducing them as input to the decoder D to obtain the representation of this z. The ith row of this plot corresponds to the traversal of the ith dimension in the latent space, whose values are traversed linearly through the different columns. In Figure 4 we can observe how the most relevant perceptual factors of the data are successfully learned by our model (e.g. the lightness, geometry, hue, glossiness, or illumination).

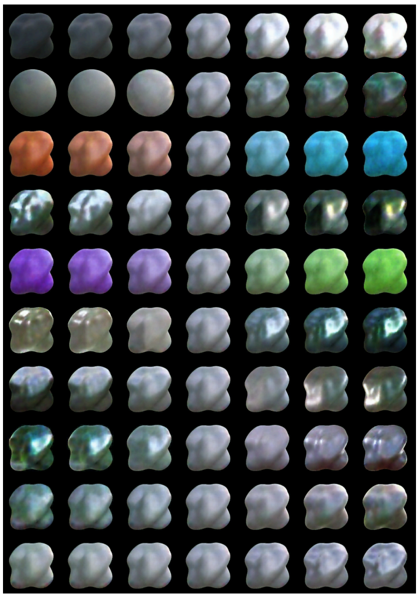

Figure 4: Prior traversals plot. Each row contains the traversal over a dimension in the latent space.

Conclusions

We propose to use an unsupervised FactorVAE model and evaluate its ability to discover disentangled and interpretable representations of material appearance in images. These representations can not only be leveraged for applications in computer graphics, such as material editing, but can also help to advance our understanding of material perception.

REFERENCES

[1]. KINGMA, Diederik P.; WELLING, Max. Auto-Encoding Variational Bayes. stat, 2014.

[2]. KIM, Hyunjik; MNIH, Andriy. Disentangling by factorising. In *International conference on machine learning*. PMLR, 2018. p. 2649-2658.

[3]. SERRANO, Ana, et al. The effect of shape and illumination on material perception: model and applications. *ACM Transactions on Graphics (TOG)*, 2021, vol. 40, no 4, p. 1-16.

[4]. STORRS, Katherine R., et al. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 2021, vol. 5, no 10, p. 1402-1417.

[5]. GUERRERO-VIU, Julia, et al. Predicting Perceived Gloss: Do Weak Labels Suffice?. In *Computer Graphics Forum*. 2024. p. e15037.