How object segmentation and perceptual grouping emerge in noisy variational autoencoders

Ben Lonnqvist, Zhengqing Wu, Michael H. Herzog

EPFL (École Polytechnique Fédérale de Lausanne), Switzerland

Humans and many newborn animals are able to effortlessly perceive objects and to segment them from other objects and the background, and a long-standing debate concerns the question of whether object segmentation is necessary for object recognition. While deep neural networks (DNNs) are state-of-the-art models of object recognition and representation, their performance in segmentation tasks is generally worse than in recognition tasks. For this reason, it is often believed that object segmentation and recognition are separate mechanisms of visual processing. Here, however, we show evidence that in variational autoencoders (VAEs), segmentation and faithful representation of data can be interlinked.

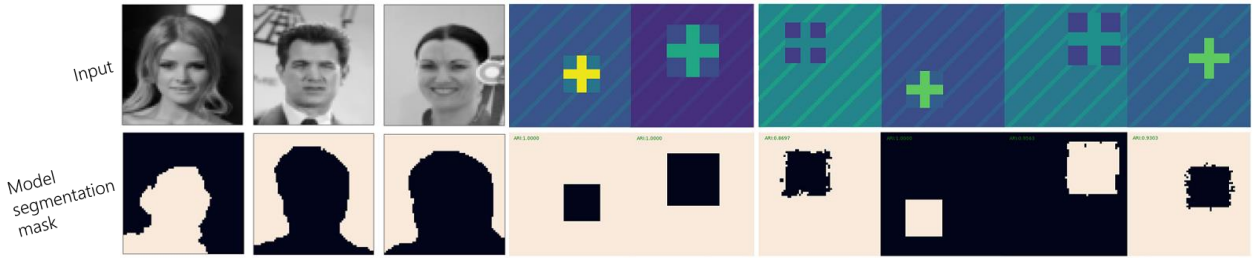

Figure 1. Model input (top) and segmentation masks (bottom) produced for CelebA faces (1-3), for Swiss flag stimuli (4-5), and for Illusory contour stimuli (6-9). The model groups objects that belong together regardless of the specific pixel values of the stimuli.

VAEs are encoder-decoder models that learn to represent independent generative factors of the data as a distribution in a very small bottleneck layer - for example, when coding for a face, VAEs may empirically code for mouths and eyes independently. Specifically, we show that VAEs can be made to segment objects without any additional finetuning or downstream training. This segmentation is achieved with a procedure that we call the latent space noise trick: by perturbing the activity of the bottleneck units with activity-independent noise, and recurrently recording and clustering decoder outputs in response to these small changes, the model is able to segment and bind separate features together.

Specifically, we trained the model with the standard (β)-VAE objective on several datasets of interest, including CelebA and a synthetic Swiss flag stimulus dataset. Once trained, we repeatedly perturb the latent vector $\boldsymbol{\mu}(\boldsymbol{x}^i)$ of the model by adding a small amount of activity-independent noise: $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}^i) = \boldsymbol{\mu}(\boldsymbol{x}^i) + \boldsymbol{\epsilon}$. This results in a model output reconstruction image $\tilde{\boldsymbol{x}}^i = Dec(\tilde{\boldsymbol{\mu}}(\boldsymbol{x}^i))$. We apply this process repeatedly, sampling a new noise sample $\boldsymbol{\epsilon}$ every iteration. The purpose of this procedure is to cause small changes in the latent space of the model, and to map those changes to the model output. The outputs are then stacked and clustered pixel-wise using hierarchical clustering, and this clustering is the segmentation mask.

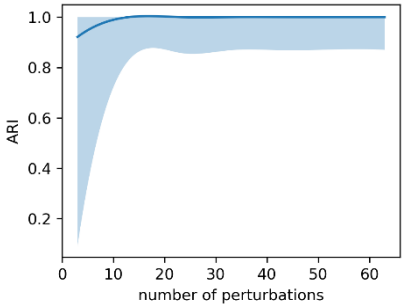

Figure 2. The model's segmentation performance (ARI) on the Swiss flag (incl. illusory contour) stimuli as a function of the number of iterations. Line shows mean performance and shaded area shows 75th and 25th percentiles respectively. The model's segmentation output quickly converges to the correct solution.

We demonstrate that VAEs can group elements in a human-like fashion, are robust to occlusions, and produce illusory contours in simple stimuli (Figure 1). Furthermore, the model generalizes to the naturalistic setting of faces, producing meaningful figure-ground segmentation without ever having been trained on segmentation. Furthermore, this segmentation takes as few as 5 iterations (Figure 2). For the first time, we show that learning to faithfully represent stimuli can be generally extended to segmentation using the same model backbone architecture without any additional training.

Funding: BL was supported by the Swiss National Science Foundation grant n. 176153 "Basics of visual processing: from elements to figures".