How to Segment Images? On Intrinsic Dimension Estimation for Patch Manifolds

Data Science Institute
COLUMBIA UNIVERSITY

Chenyu Zhang
John Wright

Data Science Capstone Project
with John Wright, EE&APAM

Patch Manifold

Many vision tasks involve *patching* images before feeding them to the model, e.g., Transformers (Dosovitskiy et al., 2020). A fundamental question is:

What is the optimal patch size?

We approach this question by investigating the patch manifold, i.e., the data manifold composed of all possible patches of the original image set.

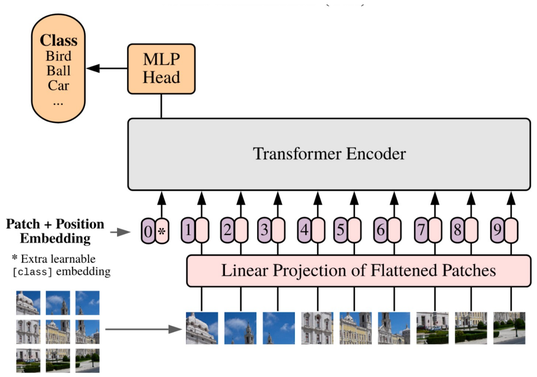

Figure 1. Vision Transformer overview (Dosovitskiy et al., 2021). Images are patched before processing.

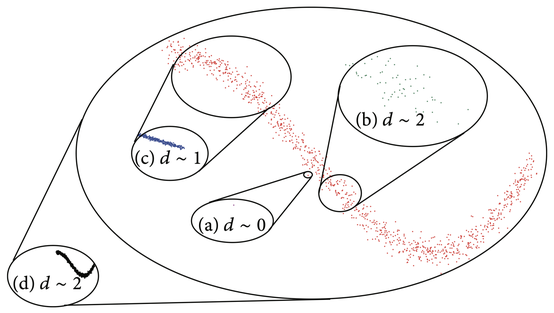

Figure 2. Illustration of different dimension estimation results at different scales (Campadelli et al., 2015).

Intrinsic Dimension Estimation

The dimension of the data manifold plays an important role in the learning performance. Similarly, we believe that the dimension of the patch manifold is also a key quantity in choosing the optimal patch size. We explored several intrinsic dimension estimation methods, including Correlation Dimension and DANCo (Campadelli et al., 2015).

Figure 3.(a). One-dimensional image set generated by rotating one image. Each image is of size 28x28.

Figure 3.(b). MNIST dataset. Each image is of size 28x28.

Figure 3.(c). CIFAR-10 dataset. Each image is of size 32x32x3.

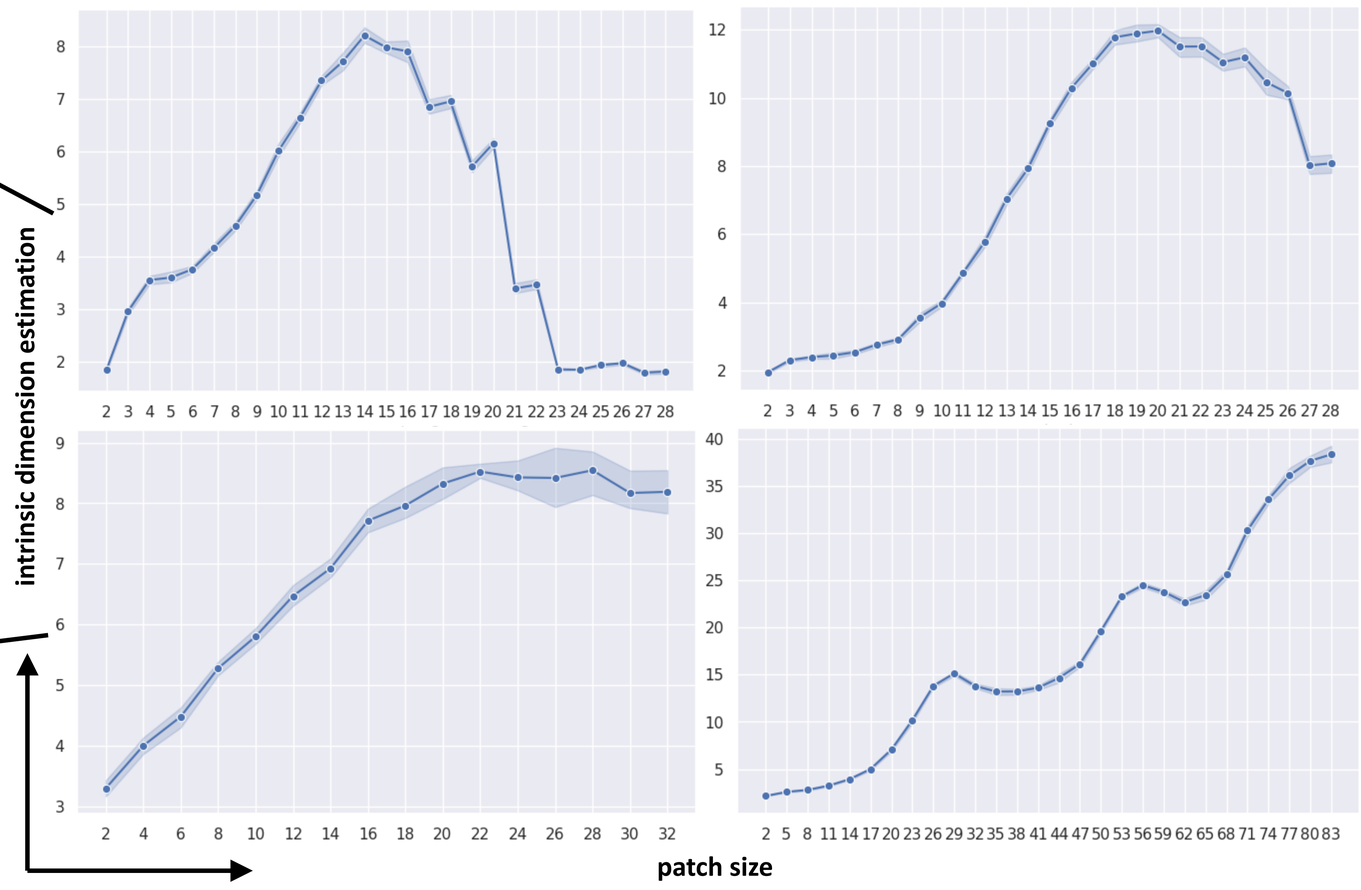

Figure 3. Intrinsic dimension estimation vs. patch size for four datasets. Each dataset contains 1000 images.

Figure 3.(d). Images with compositional structure. Each image is of size 84x84.

Next Steps

Figuring out the relationship between the patch manifold dimension and the patch size is only the first step in answering the question. Some of our next steps are: (1) derive the general law describing the relationship between the patch manifold dimension and the patch size; (2) explore the relationship between the patch manifold dimension and the learning performance of downstream tasks.

References

Dosovitskiy, A., et al., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Camastra, F. and Staiano, A., 2016. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*.