Depth estimation with Visual Transformer

Lee Ju Hwan, Dang Thanh Vu, Yu Gwang Hyun and Kim Jin Young

Abstract: In this study, we propose an approach to estimate depth information in a single image. Our built-in model utilizes Visual Transformer (ViT), a novel architecture with great success in solving a wide range of computer vision problems.

Introduction

Method

Network Architecture

This study introduces an effective yet simple framework for image retrieval. The feature descriptor is extracted from a backbone network and goes further through a projection module to become embedded vectors for retrieval. We study the behavior of the output representation space when training with a contrastive loss. In particular, how augmenting impacts the space properties and the performance of image retrieval on low-resolution inputs. Our framework is illustrated as in Figure 1. This section demonstrates our framework with three modules. A feature extractor module extracts a visual representation of a given image, a projection-head trained in a contrastive approach helps to map visual representation to an embedding space so that the similarity of samples can be calculated. Finally, we introduce an auxiliary module with classification loss and triplet loss, which significantly enhances the category retrieval's performance.

2.1. Feature extractor

We manipulate the Visual Transformer [11] model to extract a visual representation of given images. A main component of the ViT model is the self-attention encoder module, which implements Transformer architecture [10] in the most standard way. According to the original version of ViT, our feature extractor involves three main steps.

Patch embedding: we split an image into a sequence of patches and map each patch to a D dimensions embedding space. Precisely, we put an image x∈R^(C×W×H) through D 2d convolutions with the kernel size of P×P and stride of P, resulting in a feature map with the size D×N×N, then flattening the feature map into a sequence of N^2 latent vectors with a constant size of D. In the above configuration, N^2=H×W/P^2 is the number of embedded patches, where (H,W) represesents the resolution of the original image, and (P,P) is the resolution of an image patch. Apart from the embedded patches, the ViT model adds an extra learnable class embedding for classification tasks, and the results obtained from using this class token are referred to as "ViT-class" in our study. To maintain the position of the patches after flattening, we follow the standard way by adding a learnable 1D positional embedding into each patch, and this positional embedding does not share weights across patches.

Encoder: Encoder is a computational block consisting of a multi-head attention module [10] and an MLP with two consecutive linear layers. Input and output of encoder module are both embedded vector of batches. LayerNorm is applied before feeding embedded vectors into the attention module and MLP. The multi-head attention module expands the model's ability to jointly focus on different positions, thus providing different representation subspaces of pair (key K,query Q,value V) from different attention heads:

multihead=Concat(head\_1,…,head\_h ),

(1)

Where each head is a context vector from scale dot-product attention.

head\_i=softmax((Q\_i K\_i^T)/√d) V\_i, (2)

Q, K, V represents query, key and value, respectively calculated inside the transformer architecture that encodes information from the image's patches with the self-attention mechanism to mutually attend to each other. d is the dimensions of patch embeddings, which equation (2) employs d to scale the attention scores.

Experimental results

1. Dataset and preprocessing
2. Quantitative results
3. Qualitative results

Conclusions

Acknowledgments

References