

High-Quality Self-Supervised Deep Image Denoising

Seungmin Lee (profile2697@gmail.com; 2020-20866),

Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction and Motivation

Traditional denoising methods are supervised-learning bases, requiring a pair of clean and corrupted images for each input. However, collecting such pairs is difficult and often impossible. This paper proposes a novel denoising model training method that only requires unpaired collections of corrupted images to alleviate the data collection problem. The proposed method predicts the clean pixel value with Bayesian reasoning based on a given noise pixel value and neighboring pixels around the noise pixel.

2. Self-supervised Bayesian denoising with blind-spot networks

In this section, we review the self-attention and multi-head self-attention [2], which are used for the main architecture.

2.1. Self-Attention (SA) and Multi-head Self-Attention (MSA)

Self-Attention (SA) takes $\mathbf{z} \in \mathbb{R}^{N \times D}$ and transforms it using the following equations:

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = [z\mathbf{W}_q, z\mathbf{W}_k, z\mathbf{W}_v], \quad \mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{D \times H}$$

$$\mathbf{A} = \text{softmax}(\mathbf{q}\mathbf{k}^\top / \sqrt{H}) \in \mathbb{R}^{N \times N},$$

$$\text{SA}(\mathbf{z}) = \mathbf{A}\mathbf{v} \in \mathbb{R}^{N \times H}.$$

Multi-head Self-Attention (MSA) with h heads sets $H = D/h$ and runs h SAs in parallel. After then, MSA puts together the results from the SAs using concatenation and a linear projection:

$$\text{MSA}(\mathbf{z}) = \text{cat}(\text{SA}_0(z), \dots, \text{SA}_{h-1}(z))\mathbf{W}_{msa} \in \mathbb{R}^{N \times D},$$

where $\text{cat}(\cdot)$ is concatenation and $\mathbf{W}_{msa} \in \mathbb{R}^{D \times D}$.

2.2. Transformer

Transformer is constructed using a combination of multi-head self-attentions (MSA), layer normalizations (LN), and MLP layers as follows:

$$\mathbf{z}_0 = \text{cat}(\mathbf{t}_{cls}, \mathbf{t}_0, \dots, \mathbf{t}_{N-1}) + \mathbf{E}_{pos}, \quad \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

$$\mathbf{h}_{l+1} = \text{MSA}(\text{LN}(\mathbf{z}_l)) + \mathbf{z}_l, \quad l = 0 \dots L-1$$

$$\mathbf{z}_{l+1} = \text{MLP}(\text{LN}(\mathbf{h}_{l+1})) + \mathbf{h}_{l+1}, \quad l = 0 \dots L-1$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0),$$

where $\mathbf{t}_i \in \mathbb{R}^D$ and \mathbf{t}_{cls} is the i -th embedded token and CLS token proposed in BERT [1], respectively. \mathbf{E}_{pos} is a positional embedding.

3. Method: ViT and Hybrid Architecture

3.1. ViT

ViT uses $\mathbf{t}_i = \mathbf{x}_p^i \mathbf{W}_E$ where $\mathbf{x}_p^i \in \mathbb{R}^{C \cdot P^2}$ is the i -th $P \times P$ image patch from the original image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, while $\mathbf{W}_E \in \mathbb{R}^{C \cdot P^2 \times D}$ is a simple linear projector. In this case, $N = HW/P^2$.

3.2. Hybrid Architecture

Assume that $\mathbf{f} = \text{CNN}(\mathbf{x}) \in \mathbb{R}^{C_f \cdot H_f \cdot W_f}$, where CNN is a traditional convolutional neural network. Then, the proposed hybrid architecture uses each pixel of the extracted feature $\mathbf{f}_i \in \mathbb{R}^{C_f}$ as \mathbf{t}_i , where $D = C_f$ and $N = H_f W_f$.

4. Results

The transformers show relatively lower performance than convolutional neural networks when the amount of data is small. The authors claim that the lack of some useful inductive biases such as translation invariance is the cause of the performance degradation. However, if transformers are pre-trained on a large dataset, they perform better than traditional models.

5. Personal Note

This paper contributes in that it unifies scattered architectures in several fields into a single architecture (transformer). However, it still requires many computational resources and data, which is hard to use in small organizations. Moreover, this paper has some drawbacks in that this paper does not justify why this model works well. Nevertheless, we can guess that projecting image patches using the same weights (\mathbf{W}_E) implements "translation invariance," and SA's all-to-all similarity estimation helps to recover "locality."

References

- [1] J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019. 1
- [2] A. Vaswani, et al. Attention is all you need. In *NeurIPS*, 2017. 1