Train ViT on Small Dataset With Translation Perceptibility

Huan Chen, Wentao Wei, Ping Yao

Institute of Computing Technology, Chinese Academy of Sciences;
University of Chinese Academy of Sciences; Chien-Shiung Wu College, Southeast University

INTRODUCTION

- Vision Transformers (ViTs) exhibit deficiencies when trained on smaller datasets, specifically lacking locality, inductive biases, and hierarchical structure, which are inherent in convolutional approaches.
- Inspired by the translation equivariance of CNNs, we propose a novel self-supervised auxiliary task that enables ViTs to acquire translation perceptibility.
- Our method delivers competitive performance on small datasets across various resolutions without necessitating architectural modifications, and it can be seamlessly integrated with previous methods for enhanced utility.

Translation Perceptibility

Consider “x” and “y” as the input and output respectively, let “TS” denote the translation-set, and let “F” and “trans” stand for the model and translation function respectively.

- Translation Invariance means that the system produces exactly the same response (output) regardless of how its input is translated.

  \[ y = F(\text{trans}(x, TS)) = F(x) \]

- Translation Invariance means that the system produces exactly the same response (output) regardless of how its input is translated.

  \[ y = F(\text{trans}(x, TS)) = \text{trans}(F(x), TS) \]

- Translation Perceptibility means that the system can work differently in different locations, but its output changes regularly with the input.

  \[ y = F(x) \]

  \[ y_{\text{TS}} = F(\text{trans}(x, TS)) \]

  \[ TS = \text{MLP}(y, y_{\text{TS}}) \]

METHODS

Pipe: In order to guide the model in learning translation perceptibility, we first apply an arbitrary translation to the input image along any direction and generate the corresponding translation labels. Subsequently, both the original and translated images are fed into the network for processing. The output tokens are utilized for classification tasks as well as translation perception prediction tasks.

RESULTS

- Larger dimensions: Our method excels with larger input dimensions.

CONCLUSION

- We propose a self-supervised training method for Vision Transformers (ViTs) on small datasets, guiding ViTs to learn translation perceptibility.
- Our approach outperforms state-of-the-art methods on small datasets with varying resolutions, and its benefits amplify as the input size increases.
- Our approach can integrate previously advanced methods, demonstrating its extensive extensibility.

Reference