Abstract: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Koorye

2024-06-06

This is an article about the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale "speech report. In the speech, a variety of language skills and rhetorical devices are used to describe the background introduction, methods, experimental results, conclusions and other chapters in the article. The audience is shown the whole process of this paper from motivation to implementation.

The first part is the background introduction, in which the traditional language model and visual model are compared in detail. In the field of artificial intelligence, language models like GPT and BERT have revolutionized tasks like conversation, professional question answering, and machine translation by using one-dimensional word vector sequences. Instead, visual models have traditionally operated on two-dimensional pixel inputs, extracting features block by block through sliding Windows on the image. Based on the above, the main theme of this talk is: Can text processing methods in language models inspire improvements in visual models, potentially ushering in a new era of visual tasks?

Next comes the method section, where the authors propose a novel way to segment images into blocks and arrange them into a one-dimensional sequence, similar to text, which is then processed using the same structure as a language model. The model consists of multiple stacked attention modules, each of which calculates the similarity between pairs of image blocks within the sequence. These similarities inform the weighted fusion process that integrates features in the sequence. After being processed by multiple layers of concern, the output is fed to a classifier or detector for visual tasks such as image classification and object detection. In this chapter, the speech firstly explains the transformation process from image to sequence through the combination of language and action, and through action and painting. Metaphorically, the different components of the model are then compared to specific parts of the human body - for example, equating feature extraction with the eye and feature interaction with neurons - so that the image processing process in this paper can be understood even if you are not a specialist in the field.

In the experimental part, the thesis firstly compares the effect of this method with that of traditional visual processing method. With the increase of training data, the author's model outperforms traditional visual models in performance and achieves the highest accuracy in image classification tasks. Compared to traditional models, the proposed model requires a smaller size to achieve the same performance, underscoring the significant improvements in various tasks and datasets presented in this paper. In addition, the presentation also metaphorically compares the attention mechanism in the model to the human attention to the prominent foreground area, explaining the model's ability to capture information-rich visual areas, which provides clues to the success of this paper.

The conclusion can be summarized as follows: Inspired by the language model, the author processes the image into a sequence similar to the text and uses the same structural method to process the visual information. This model is superior to the traditional vision model in both accuracy and computational efficiency, which lays a new foundation for the development of this model. After that, the presentation summarized the limitations of this paper and the future development, due to the limitations of hardware resources, the full potential of this new model has not yet been realized. Future work includes experimenting with larger data sets and increasing the size of the model. The performance of this innovative model in a variety of downstream tasks, such as object detection and human pose recognition, remains to be fully evaluated.

This talk introduces a new model that bridges the gap between linguistic and visual models. This finding points to a promising direction for future research on vision tasks, with the potential to redefine standards for model performance and efficiency. My contributions to the speech include: speech preparation, making beamer, recording video of the speech, live speech, writing abstract, etc. I hope this speech can enlighten the audience.