Programming Assignment 2: PoseRAC

Bingnan Li
2020533092
libn@shanghaitech.edu.cn

Abstract

In this assignment, I explored the state-of-the-art repeat action counting algorithm **PoseRAC** *and tried to reproduce the results in the original paper. Then I proved the necessity of Transformer encoder by canceling the encoder part and using Fully-Connected Layer only. Moreover, I also tried to explore the relationship between counting performance and the number of encoder layers together with number of heads in the multi-head attention module. The results show that the performance of PoseRAC barely improves when the number of encoder layers increases from 1 to 8, but the model will crush when the number of encoder layers is larger than 8 in my training setting. Besides, the performance of PoseRAC is not sensitive to the number of heads in the multi-head attention module. Finally, I replaced the triplet margin loss with* **contrastive loss** *and* **circle loss***, the results show that circle loss significantly boosts the model within 20 epoch, the evaluation metrics MAE and OBO improved from 0.2540 and 0.5395 to 0.2083 and 0.6053.*

1. Introduction

PoseRAC is the state-of-the-art repeat action counting algorithm proposed by [1]. This model achieves tremendous success in the repeat action counting task and improves the performance of the previous state-of-the-art model by a large margin. The novelty of PoseRAC is the new annotation method. The traditional method will annotate the start and the end frame of an action, models are forced the regress the locations or indices of an action. However, PoseRAC turns the annotation into two salient frames which indicates the most representative points of an action. Then PoseRAC utilizes a keypoint extractor to transform the salient frames into a series of keypoint with 3D coordinates. This operation enormously reduces the amount of data need to process and improves the effectiveness of information because it only uses transformer encoder layer to get the embedding information and a single FC layer to classify the embedding.

Model	MAE	OBO
PoseRAC	0.2540	0.5395
w/o Transformer	0.9928	0.0263

Table 1. The performance of PoseRAC with and without transformer encoder.

2. Necessity of Transformer Encoder

In this section, I will prove the necessity of Transformer encoder by canceling the encoder part and using Fully-Connected Layer only. By the original implementation, the input of the model is a series of keypoint with 3D coordinates, the output of the model is the score of different salient frame of different actions. The keypoint data is fed into a transformer encoder layer and a fully-connected layer to get the final output. Since the transformer does not change the dimension of the input, I can directly cancel the transformer encoder layer and use a fully-connected layer to get the final output to explor how essential transformer is to the model.

The comparison of the performance of PoseRAC with and without transformer encoder is shown in Table 1. This result shows that the performance of PoseRAC is significantly degraded when the transformer encoder is canceled. Even the training loss (Binary Cross Entropy) drops to the same level of that of the original model, the evaluation metrics MAE and OBO are still much worse than the original model.

This experiment proves that directly classifying the keypoint data with 3D coordinates (normalized to $[0, 1]$) cannot fit the test data distribution and the transformer encoder embeds the keypoint information into a feature space that are highly separable and truly represents the common features for the same action.

Training Setting:

- lr: 0.000025
- α: 0.01
- epoch: 20

3. Impact on the Number of Encoder Layers and Number of Heads

In this section, I will explore the relationship between counting performance and the number of encoder layers together with number of heads in the multi-head attention module. In the original implementation, the number of encoder layers is 6 and the number of heads is 9. Intuitively, the more encoder layers and heads, the more information the model can capture. So, I set the number of encoder layers to be (1, 2, 4, 6, 8, 10, 12) and the number of heads to be (1, 3, 9, 11).

References

[1] Ziyu Yao, Xuxin Cheng, and Yuexian Zou. Poserac: Pose saliency transformer for repetitive action counting. *arXiv preprint arXiv:2303.08450*, 2023. 1