TOWARDS DISCRIMINATIVE SEMANTIC RELATIONSHIP FOR FINE-GRAINED CROWD COUNTING

Anonymous ICME submission

7. APPENDIX

7.1. Comparison between the Fixed and Trainable Word Vectors

In Table 5, we compare the influence of the fixed and trainable word vectors in WVM. It can be observed that the fixed word vectors achieve better CMAE in Standing/Sitting subtask but are outperformed by trainable word vectors in other three subtasks. In our opinion, the initial word vectors in this subtask can model the semantic relationship well and thus need no further training. Thus, in our experiments, the word vectors are fixed for Standing/Sitting subtask and are trainable for the rest three subtasks.

task	method	**CMAE**
Standing/Sitting	fixed	**5.87**
	trainable	6.39
Waiting/Not waiting	fixed	2.74
	trainable	**2.48**
Towards/Away	fixed	2.02
	trainable	**1.82**
Violent/Non-violent	fixed	4.02
	trainable	**3.51**

Table 5. Comparison between the fixed and trainable word vectors.

7.2. Visualization for Other Two Subtasks

To further understand the effectiveness of our proposed approach, we provide more visualization results here. In all visualizations below, the left column shows the visualization of the baseline and the right column shows ours. The text in the lower left corner of each picture indicates the category.

In Fig. 6, we show the visualization of fine-grained density maps generated by the baseline and our proposed approach for Violent/Non-violent subtask. In the first row, one fighting boy (in the blue box) is misclassified by the baseline as non-violent, while ours can correctly classify him as violent. In the second row, the hand on the back and the food on the table (in the yellow boxes) somehow mislead the baseline, while ours is less affected by the background thanks to the proposed AKM. In the third row, we provide an ordinary example to show that our semantic segmentations are more refined and concentrated on the corresponding people category, with the aid of the proposed discriminative semantic relationship modeling.

Waiting/Not-waiting subtask is more challenging because whether a person is waiting for a bus or not is mainly decided by his context (other people and bus stop), not his appearance, which depends more on the semantic relationship in the image. In Fig. 7, we show the visualization of fine-grained density maps generated by the baseline and our proposed approach for Waiting/Not-waiting subtask. To make it clearer, we use arrows to indicate the enlarged original pictures of the regions of interest.

In the first row, the baseline misclassifies the waiting people on the right of the picture where there is obviously a bus stop. In the second row, the baseline judges the woman in the center of the enlarged original picture as not waiting. But she is surrounded by the waiting people and is actually waiting for the bus. Our proposed approach obtains the correct results in these examples. In the enlarged original picture of the third row, two women are on either side of the waiting people. They are ignored by the baseline because they are not closely connected to other waiting people, while ours can correctly classify them by modeling their semantic relationship with others. In the fourth and fifth rows, the arrow-indicated waiting people are considered as not waiting by the baseline. By better exploring the semantic relationship in the image, our approach largely avoids the above situations and achieves clearly better performance in this fine-grained crowd counting task.

Fig. 6. The visualization of fine-grained density maps generated by the baseline and our proposed approach for Violent/Non-violent subtask.

Fig. 7. The visualization of fine-grained density maps generated by the baseline and our proposed approach for Waiting/Not-waiting subtask.