Joint Event Recognition and Image Curation (Supplementary Material)

Anonymous ECCV submission

Paper ID 1644

1 Performance - Iteration number

In the main paper, we described the iterative curation-recognition procedure which iteratively updates album-wise event type prediction and image-wise importance score prediction. In case the procedure does not converge and oscillate, the final event type / image importance prediction is averaged over final 3 iterations. There are two hyperparameters for the iterative procedure: $\theta = (m, \alpha)$. Here, m is the threshold of minimum fraction of the maximum probability to consider, which eliminates the event types with low confidence for image importance prediction; α is the emphasis we give to image importance score for event type prediction.

The hyperparameters are decided with a validation set. In this section, we analyze the algorithm performance with respect to the iteration number on ML-CUFED Dataset.

1.1 Event Recognition Performance - Iteration Number

There is more than one possible choice for θ from the event recognition result of validation set, and if we look into the trend of recognition accuracy with iteration number, there are several different types, as shown in Figure 1. To show the different trends more clearly, Figure 2 provides individually each plot for different θ.

We can see that it takes different steps to converge for different choices of θ. For Figure 2(a), $m = 0.9$, and this means only event types with very high prediction confidence are considered to contribute to the image importance prediction. It takes two steps to converge. In contrast, for Figure 2(c), $m = 0.1$, and this means that more events can contribute to the importance prediction. This results in the slower changes of image importance score prediction and event type prediction over iterations, and as shown in Figure 2(c), it takes 5 iterations to converge. For Figure 2(b), $m = 0.3$, and it takes 3 iterations to converge, which is in the middle of Figure 2(a) and (c).

Figure 2(d) also shows the case where the iterative algorithm does not converge, and the final recognition result is produced by averaging the final 3 iterations.

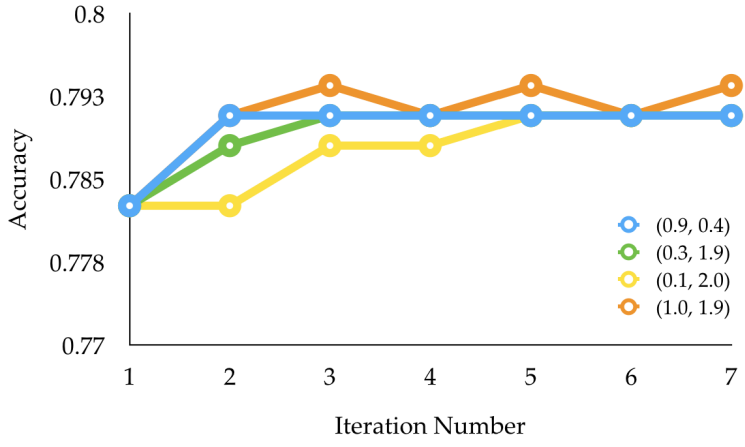

Fig. 1: Album-wise event recognition accuracy v.s. iteration number for four different choices of hyperparameters (m, α) of the iterative curation-recognition procedure.

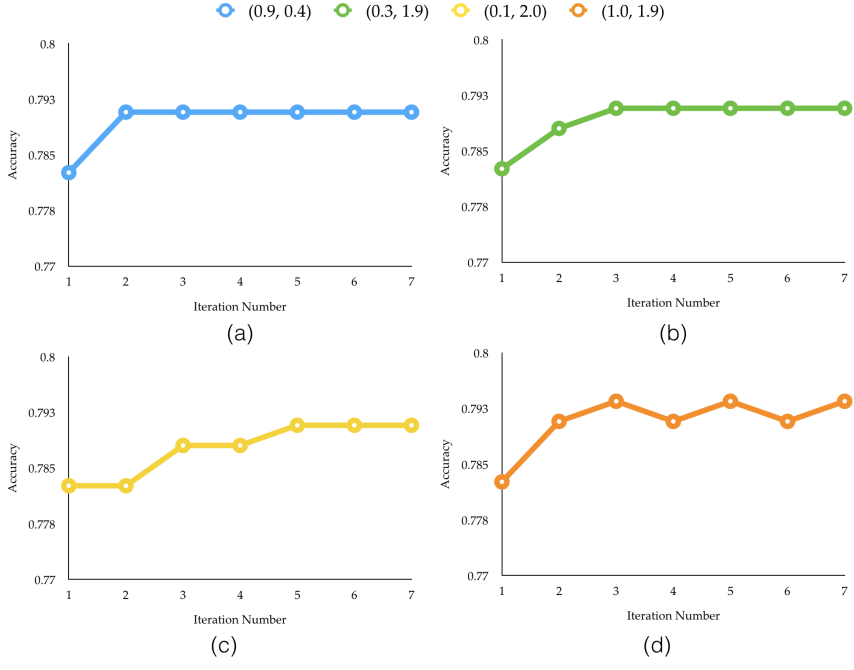

Fig. 2: (Separately shown) Album-wise event recognition accuracy v.s. iteration number for four different choices of hyperparameters (m, α) of the iterative curation-recognition procedure.

1.2 Image Importance Prediction Performance - Iteration Number

Starting from the possible choices of θ we get in Section 1.1, we can further filter the choice with the image importance prediction result on the validation set. The image importance score performance (measured by MAP@10% with iteration number) for the final choice of θ is shown in Figure 3. Note that the iterative updating algorithm is initialized with equal image importance score for every image. Therefore, first iteration importance score is calculated from the event recognition result with 78.2% accuracy, while first iteration event recognition is calculated from equal importance score, which is obviously not a good image score prediction. Therefore, we expect a greater gain after iterative updates for event recognition than image importance prediction. As shown, the performance converges after second iteration.

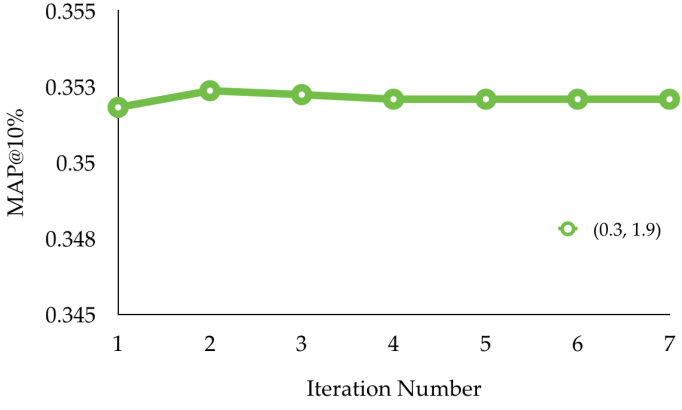

Fig. 3: For final choice of $\theta = (m, \alpha) = (0.3, 1.9)$, the image importance prediction accuracy v.s. iteration number.

We can see that the change of MAP score over iterations are no larger than 0.2%, which is not significant.

2 Qualitative results

3 Long Short-Term Memory(LSTM) Network

The architecture of the CNN-LSTM network for album-wise event type prediction is shown in Figure 4. The images of an album is first fed into the trained CNN for single images, and the 7th fully connected layer feature is extracted for each input image. The dimensionality of the FC7 features are then reduced to 128 with PCA. Then the sequence of features is fed into the LSTM unit to predict the event type of the album for each time frame. The predictions are fed into a mean pooling layer for the final prediction.

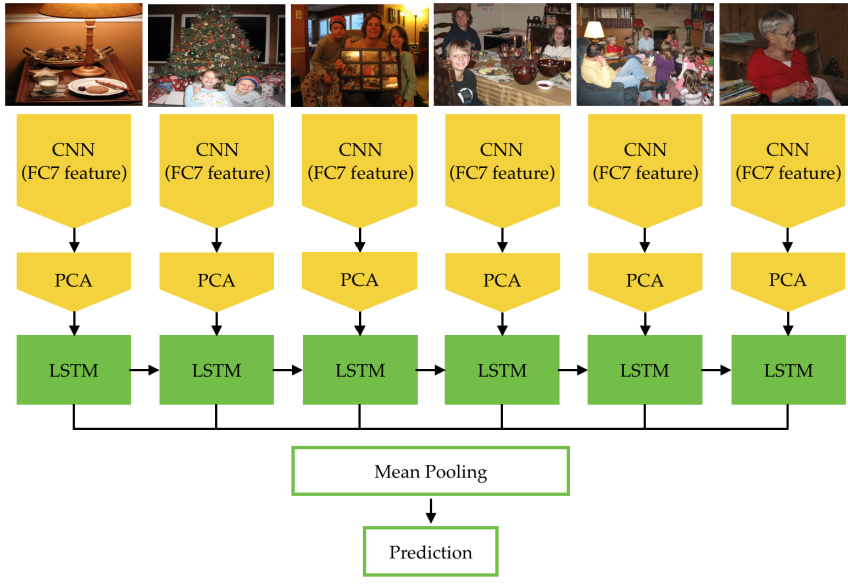

Fig. 4: Architecture of the CNN-LSTM network