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 last 3 iterations. There are two hyperparameters for the iterative procedure: $\theta = (m, \alpha)$. Here, m is the threshold for the 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 put on the 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 the recognition accuracy over the iteration number, there are several different types, as shown in Figure 1. To show the different trends more clearly, Figure 2 illustrates individual 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 only 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 last 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. This is average MAP over the entire test set. 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.

2 Qualitative results

In this section, we show more examples of the qualitative result from our algorithm.

In Figure 4, four albums in ML-CUFED Dataset are shown. For the examples shown here, the album-wise event type prediction is wrong with CNN-recognition method (with only a CNN to predict event type of single images, and averaging the prediction from every image in an album for final prediction), but with the proposed CNN-LSTM-Iterative algorithm, the event type prediction is correct. Also, we can see the ground-truth and predicted importance ranking of the images in each album. Only a fraction of images in the albums are shown here due to limited space. We deliberately choose both images with high ground-truth

importance and low ground-truth importance for each album to show the overall quality of the albums.

In the first example of a Cruise Trip album, there are many images of the iceberg and the sea like the last image shown here. If only CNN-recognition is used, and the prediction is produced by averaging the prediction of every image in the album, and the album is recognized as Beach Trip. However, after we assign different importance score to images, as shown the ranking of images, this album is correctly recognized as Cruise Trip. This indicates the advantage of the joint recognition-curation algorithm. In other words, our full method is able to rank important images (which are more indicative of event types) on the top.

Fig. 4: Examples of recognition-curation result from ML-CUFED Dataset. With only CNN-recognition described in the main paper, the album-wise event type prediction is wrong, as shown in the first label in the right to each album; after CNN-LSTM-Iterative algorithm, the album-wise event type prediction is correct, as shown in the second label in the right. Below the images, we show the ground-truth ranking of each image in the album and the predicted importance ranking of each image in the first and second row respectively.

In Figure 5, more examples with correct event type prediction in ML-CUFED Dataset are shown. We can also see how the images are ranked with predicted importance by the algorithm.

Fig. 5: More examples of recognition-curation result from ML-CUFED Dataset. The event types of albums are correctly recognized, as shown in the right of each album. Below each album, we show the ground-truth ranking and the predicted importance ranking for each image in the first and second row respectively.

In Figure 6, we show some examples with wrong event type prediction in ML-CUFED Dataset. The ground-truth event type of the three example albums are book signing event, ball, and graduation party respectively.

Fig. 6: Examples of recognition-curation result from ML-CUFED Dataset, whose event types are predicted incorrectly. The predicted event type and the ground-truth event type are shown in the right of each album. The ground-truth event type is shown in the parenthesis.

3 Long Short-Term Memory(LSTM) Network

The architecture of the CNN-LSTM network for album-wise event type prediction is shown in Figure 7. The images of an album are 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 is then reduced to 128 with PCA, and the sequence of compressed 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. 7: Architecture of the CNN-LSTM network