Audio Visual Attribute Discovery for Fine-Grained Object Recognition (AAAI 2018 Conference)

Rishabh Ramteke

Problem Statement and Author’s Approach

The key challenge of fine-grained recognition is how to learn the discriminative feature to tell the subtle visual differences between the subcategories. Most of the fine-grained recognition algorithms require visual supervision to learn the discriminative feature representations which is time consuming and requires professional knowledge to obtain the accuracy annotations. Also, the learned representation is still far from solving the fine-grained problems due to the limitations of supervised labels. For example, the variations of visual appearances would cause the region based representation ambiguous, and then weaken the discrimination of features. In addition, the textual descriptions are subjective which may not be consistent across persons. This would let the learned representation incorrect and limit the performance of finegrained recognition.

To solve this, the authors have introduced Audio Visual Attributes feature which consists of the encoder module and the attribute discovery module, to encode the image and audio into vectors and learn the correlations between audio and images, respectively with the softmax loss function at the end. They have used feed forward CNN for encoder module and attention framework (based on LSTM) for attribute discovery module. This architecture is used by the authors to solve these challenges: (i) how to encode the audios from the untrimmed videos.(ii) the way of discovering the correlations between the audio and the visual images (iii) how to efficiently infer the defined audio visual attribute from the images and train the discriminative classifiers.

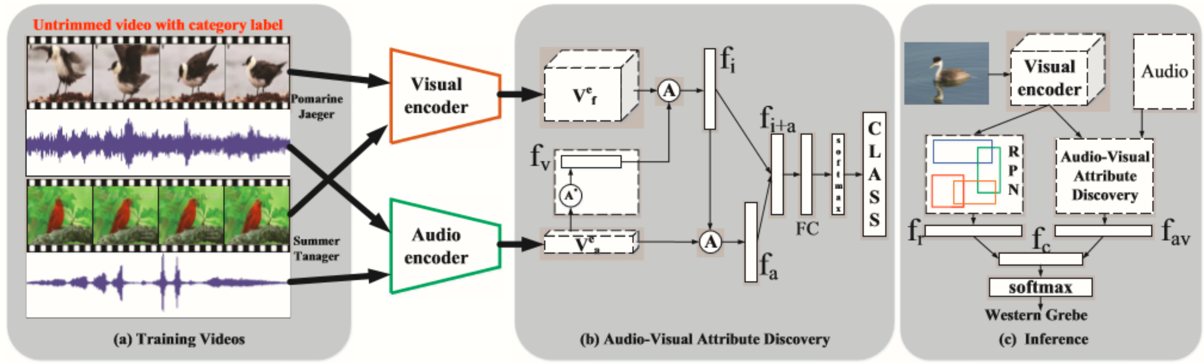

Figure: Overview flowchart of the author’s proposed architecture

Explanation of Flowchart : Given the untrimmed videos with the category label, they extracted the visual and audio information which are fed into the feature encoder module to achieve the representations. (b) After that, they trained the audio visual attribute discovery module based on the recurrent neural network. The loss function of this module is employed the softmax with the category label. (c) In the step of the inference, they extracted the proposals with region proposal network (RPN), and then combine the visual features and the audio visual attribute to generate the image representation. Finally, a classifier is learned based on the representations to predict the image category.

Strengths

The Advantage of their proposed feature is that it is easily achieved without the human intervention,which is more objective than these textual descriptions. Also, the audio visual features is robust which can encode the global feature of the object.

Drawbacks

In some examples, attribute discovery module fails. It can be observed that without sufficient training samples, the audio visual attributes model is easily degenerated. They should have designed an end-to-end training neural network instead of with two stages. The model's working is only shown on bird species dataset in which it performs really good but one dataset doesn't seem sufficient to convince that their algorithm is better than other baseline methods.

Future work suggestions

In the future, they could focus on how to fuse the multiple supervisions to learn the fine-grained object classifier and how to improve the discrimination of image representation via merging the multi-view features. They could work on how to develop the hierarchical structure representation. Also,they could treat the audio information as the supervision to learn the features from the visual images.

They should see if their model works on other datasets as well. And as I suggested in the previous section, they should have designed an end-to-end training neural network instead of with two stages