In our project Echo, animal sounds are transformed into Melspectrogram and trained through CNN. Therefore, I have an idea whether we can use an efficient *Image Multi-label Classification* to solve the problem.

HCP (hypothesis-CNN-Pooling): A Flexible CNN Framework for Multi-Label Image Classification (https://ieeexplore.ieee.org/document/7305792)

The model construction:

(1) BING (binarized normed gradients) extracts proposals for some objects, and then use HS (hypothesis selection) method selecting some hypotheses from these proposals.

(2) Input hypothesis into the shared CNN network, the network is pre-trained with single-label data. For each hypothesis, a C-dimension prediction result was output.

(3) Fine-tuning parameters, training for multi-label data. cross-hypothesis max-pooling is adopted to generate the final prediction results.

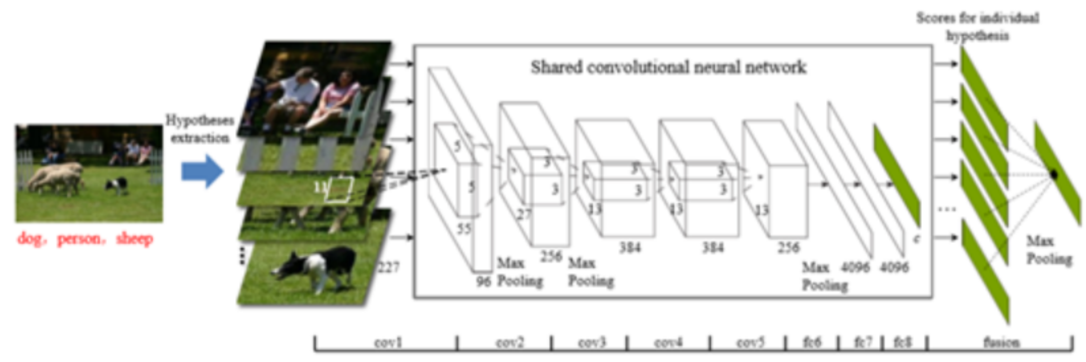


Fig. 2. An illustration of the infrastructure of the proposed HCP. For a given multi-label image, a set of input hypotheses to the shared CNN is selected based on the proposals generated by the state-of-the-art objectness detection techniques, e.g., BING [8]. The shared CNN has a similar network structure to [24] except for the layer fc8, where c is the category number of the target multi-label dataset. We feed the selected hypotheses into the shared CNN and fuse the outputs into a c-dimensional prediction vector with cross-hypothesis max-pooling operation. The shared CNN is firstly pre-trained on the single-label image dataset, e.g., ImageNet and then fine-tuned with the multi-label images based on the squared loss function. Finally, we retrain the whole HCP to further fine-tune the parameters for multi-label image classification.

How to do hypothesis selection:

As shown in the picture below. There are m groups, and in each group, the first k hypothesis is taken according to the predicted score of BING. So, there are m*k hypotheses.

(This is a report about BING: https://ieeexplore.ieee.org/document/7428721)

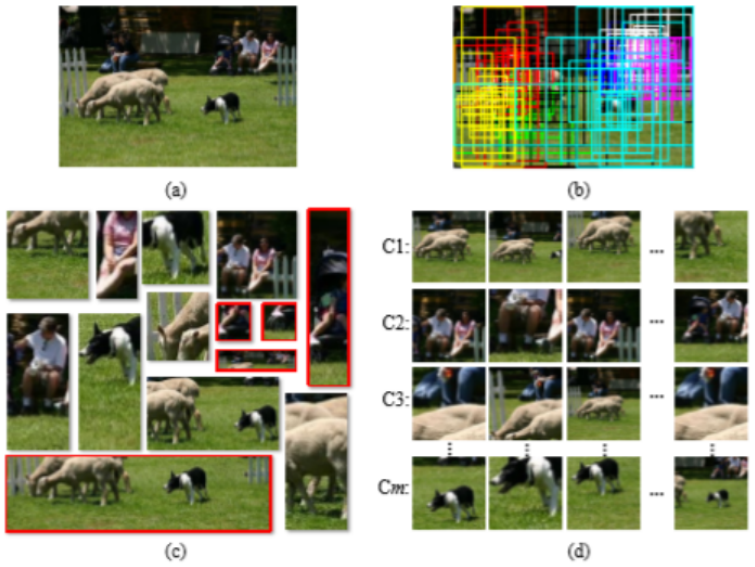
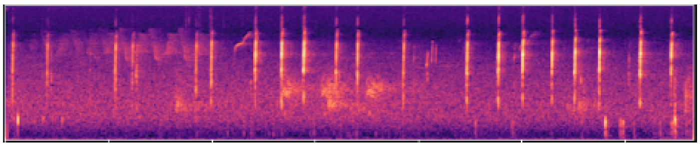
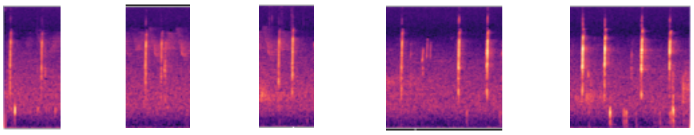


Fig. 3. (a) Source image. (b) Hypothesis bounding boxes generated by BING. Different colors indicate different clusters, which are produced by normalized cut. (c) Hypotheses directly generated by the bounding boxes. (d) Hypotheses generated by the proposed HS method.

For example, this is a Melspectrogram have multi-labels:





Initialize HCP:

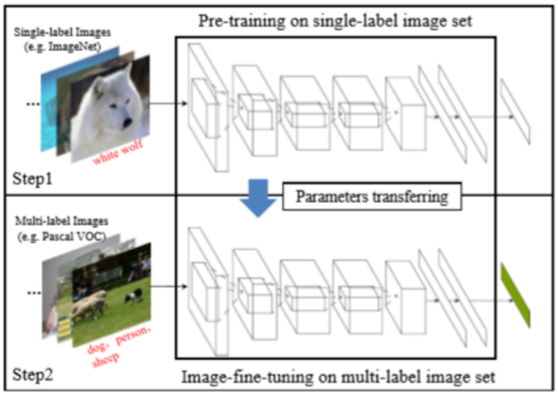


Fig. 4. The initialization of HCP is divided into two steps. The shared CNN is first pre-trained on a single-label image set, e.g., ImageNet and then fine-tuned on the target multi-label image set using the entire image as input. Parameters pre-trained on ImageNet are directly transferred for fine-tuning except for the last fully-connected layer, since the category numbers between these two datasets are different.

Hypothesis fine tuning:

H-FT is the method for cross-hypothesis max-pooling. See the following figure, combines the predicted results of each hypothesis and eliminates the interference of noise at the same time.

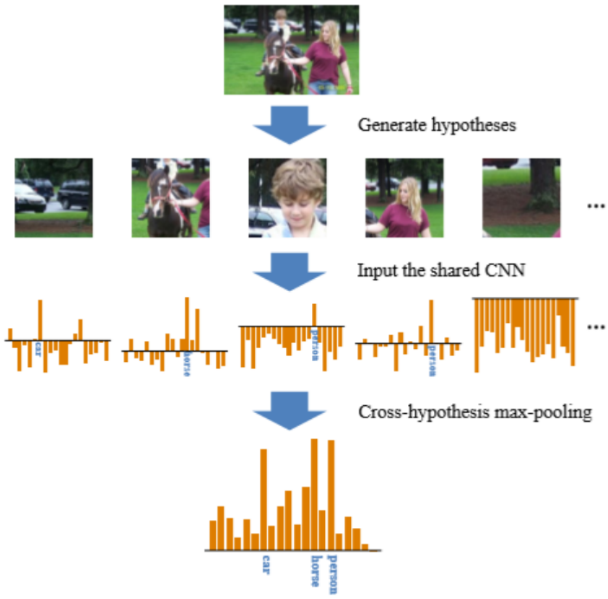
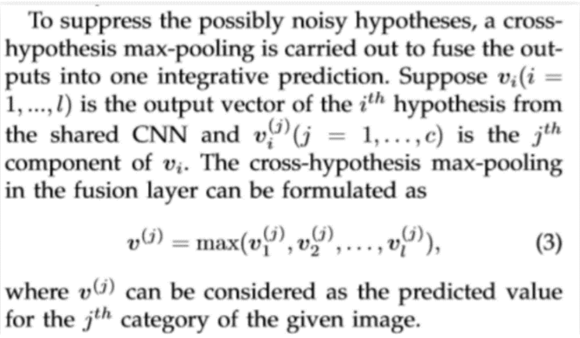


Fig. 5. An illustration of the proposed HCP for a VOC 2007 test image. The second row indicates the generated hypotheses. The third row indicate the predicted results for the input hypotheses. The last row is predicted result for the test image after cross-hypothesis max-pooling operation.



To suppress the possibly noisy hypotheses, a cross-hypothesis max-pooling is carried out to fuse the outputs into one integrative prediction. Suppose $v_i(i = 1, ..., l)$ is the output vector of the i^{th} hypothesis from the shared CNN and $v_i^{(j)}(j = 1, \ldots, c)$ is the j^{th} component of v_i. The cross-hypothesis max-pooling in the fusion layer can be formulated as

$$v^{(j)} = \max(v_1^{(j)}, v_2^{(j)}, \ldots, v_l^{(j)}), \qquad (3)$$

where $v^{(j)}$ can be considered as the predicted value for the j^{th} category of the given image.