Description of the algorithm

(1) Contact

- Name: Satoshi Kondo
- Affiliation: Konica Minolta, Inc.
- E-mail: satoshi.kondo.jp@gmail.com

(2) Team name: KM

(3) Overall structure of our algorithm

Summary

We use 50 layer ResNeXt [1] deep learning network with Squeeze-and-Excitation block [2]. The network is pre-trained with ImageNet dataset. We add two fully connected layers on top of the ResNeXt network. The parameters of the base network, i.e. 50 layer ResNeXt, are fixed and the additional layers are only trained. The network is trained frame by frame and we do not use temporal information.

We trained two models. These models have the same network structure as mentioned above. We trained these models by using different configuration in the training phase.

The common training configuration is as follows. An input image is resized to 640 x 480 pixels. We use stochastic gradient descent for the optimization. The hyper-parameters in the optimization are that the initial learning rate is 0.01 and the momentum is 0.9. We decay the learning rate with a cosine annealing for each epoch. The mini-batch size is 15 and we run 100 epochs. The loss function is sigmoid cross entropy. The loss function is weighted depending on the probability of the classes (present or not present) for each tool as $w_{class} = 1 / \ln(1.02+p_{class})$ [3], since the classes for each tool are highly imbalanced.

The different configuration for two models is as followings. In the first model, we select frames every 10 frames from the training dataset. We do not use data augmentation. In the second model, we select frames every 20 frames including tools #1, #2, #3, #5, #6, #10, #11, #16, #17, #19 and #21 in addition to the frames selected for the first model. And data augmentation is applied on the fly during the training. The type of the augmentation is translation, rotation, resizing, flipping and contrast.

At the inference time, the probabilities presenting each tool are obtained from the output of the fully connected layer on top of the base network. We have two sets of the probabilities from two models, and we take the averaged values. And we apply post filtering to averaged probabilities for each tool in temporal direction and get the final probabilities.

References

[1] S. Xie, “Aggregated Residual Transformations for Deep Neural Networks,” CVPR 2017.
[2] J. Hu, et al., “Squeeze-and-Excitation Networks,” CVPR2018.
[3] A. Paszke, “ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation,” arXiv:1606.02147.

(4) We do not use additional training data. But we use a model pretrained by using ImageNet dataset.

(5) Predictions for a given frame solely rely on that frame. But we apply post filtering to probabilities for each tool in temporal direction.

(6) Predictions for the microscope video frames solely rely on microscope video frames. We do not use tray video frames.

(7) In prediction, number of frames processed per second is about 16 frames including reading image data from disk and preprocessing for one model. We use two models, so number of frames processed per second is about 8 frames when we use one GPU. We use Nvidia GTX 1080Ti GPU and mini-batch size is 100.

(8) The algorithm has not been tested on other datasets.