Deep learning based surgical tool presence detection in cataract surgery

Muneer Ahmad Dedmari, Sailesh Conjeti
Technical University of Munich
dedmari_muneer@tum.de,sailesh.conjeti@tum.de

ABSTRACT

This document is written in support for submitting cataract surgery tool presence prediction results to CATARACTS challenge[2]. This work is still in experimental phase, hence detailed documentation is not provided.

1 PURPOSE

A fully automated surgical tool presence detection algorithm,namely, TUMCTNet, is proposed for cataract surgery video streams.

2 ALGORITHM OVERVIEW

Deep learning[6], is one of the most successful machine learning method, which allows deep neural networks to recognize the representations from raw data for specific tasks such as classification[5] and detection [4]. Recent studies such as[7] and [1] have demonstrated that deep learning also performs well for multi-label classification problem.

In this algorithm we formulate a deep learning based multi-label classification method for surgical tool detection in cataract surgery videos.

Our proposed CNN architecture is based on Inception-v4[3] network. Last layer of proposed network Is fully connected layer, consisting of 22 units for twenty one tools and an additional "No Tool" label, which is added to the ground truth and represents that none of the given tools are present in the image. Convolutional layers are initialized with pretrained ImageNet weights while as fully connected layers are initialized with random weights.

As more than one surgical tool can be present in a video frame, this motivates us to treat the problem as a multi-label classification, where co-occurrence entries are also considered valid and not penalized during classification. In this algorithm, we considered converting multi-label classification problem into several independent binary classification problems. Also, training data is highly unbalanced due to the imbalance associated with the tool usage during surgeries. To overcome this issue, we tried to balance dataset by using under-sampling techniques[8] and we used weighted sigmoid cross-entropy as the loss function.

Out of 494868 video frames (training data) we have extracted two stratified data-sets, each comprises of around 14000 frames. As this is multi-label problem, we have also considered co-relation of tools. Two models(based on above discussed architecture) were trained independently using two stratified data-sets respectively. To obtain final result we have used ensemble of these two networks.

For detecting number of tools(cardinality) present in each frame of given video, we are training third model based on above discussed network. Last fully connected layer of this architecture has 3 outputs. Each output represents number of tools present in given frame. Weighted soft-max cross entropy has been used as the loss function. Ensembled results are modified based on the prediction of number of tools.

Data is normalized by subtracting mean from every image. Further more, we perform real-time data augmentation (flipping, scaling, cropping) to avoid over-fitting the model. Each frame is resized to 640x360. Finally, the network is trained using stochastic gradient descent with learning rate of 0.001 and momentum of 0.9. For training network that predicts number of tools, we are using Adam optimizer with learning rate of 0.0001 and momentum of 0.9. On average, 17 frames per second can be processed during inference using GTX TITAN X.

3 RELATION TO PREVIOUS SUBMISSIONS

Input size of network has been changed to 640x360. For training, number of samples has been increased.

REFERENCES

[1] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen. 2015. Polyphonic sound event detection using multi label deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN.2015.7280624

[2] CATARACTS. 2017. Challenge on Automatic Tool Annotation for cataRACT Surgery. https://cataracts.grand-challenge.org//. (2017).

[3] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261. (2016).

[4] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* abs/1311.2524 (2013). http://arxiv.org/abs/1311.2524

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[6] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (2015), 436–444.

[7] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. 2016. Meka: A Multi-label/Multi-target Extension to Weka. *J. Mach. Learn. Res.* 17, 1 (Jan. 2016), 667–671. http://dl.acm.org/citation.cfm?id=2946645.2946666

[8] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. *On the Stratification of Multi-label Data.* Springer Berlin Heidelberg, Berlin, Heidelberg, 145–158. https://doi.org/10.1007/978-3-642-23808-6_10