Deep Learning for Automatic Cataract Surgery Tool Annotation

Navdeep Dahiya[*], Senthil Ramamurthy[†]

November 29, 2017

Abstract

Cataract surgery is one of the most common type of surgery in the world. Automatic annotation of which tools are being used in cataract surgery videos has important applications in workflow analysis and training. This article presents a deep learning based algorithm that is being trained and tested on the expert annotated data provided by the organizers of CATARACTS Challenege[1]. Machine Learning/Deep learning has been very successful in large scale image recognition tasks in recent years. It is finding new applications in autonomous cars, virtual reality, face recognition amongst many more. Medical image processing field has great many challenging problems where computer vision, image processing and machine learning techniques can be applied to help solve problems and develop new tools to help improve the clinical workflow and improve patient care and medical training.

1 Description of the proposed algorithm

Since 2010, the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [2]is a competition where research teams evaluate their algorithms on a given data set, and compete to achieve accurate visual recognition tasks. One of the main challenges is to identify a never before seen image to belong to one of several thousand of classes. Over the last few years many algorithms have been proposed which have achieved increasingly higher accuracy over the years. This has also led to the development on transfer learning where some kind mulit-layer convolutional neural network can be trained to detect new classes of images which have never been seen before. We have decided to use such an approach as starting point for our algorithm.

We used a Google Inception V3 [3] model that had been pretrained on the imagenet dataset and retrained it to detect 22 different tools present in cataract surgery training images provided by CATARACTS challenge. The training

[*]ndahiya3@gatech.edu
[†]senthil.ramamurthy@gatech.edu

videos had 21 different tools being used where one or more tools could be used simultaneously or no tool could be used in any given frame. We created an extra class named Tool 22 which means no tool is being used in a particular frame.

During the re-training process we extracted and separated frames from the surgery videos corresponding to each tool being used as annotated in the training-labels file. In case no tool was being used we labeled that as Tool 22. In this iteration of training we only used frames when a tool was being used only by itself rather than simultaneously with other tool. We then used Google open-source Tensorflow tool kit to retrain the Inception V3 model with a learning rate of 0.01 for several thousand iterations with cross-entropy loss. The image frames were resized to 299x299 for training and inference.

Once the new model was trained, we ran inference on all the test videos for the presence of 22 tools. While writing the prediction files in the required format, we simply ignore the Tool 22 prediction. Simply the presence of Tool 22 class during training and inference lowers the probability of other 21 tools. We achieved a fairly high accuracy of 94% on 20% of training frames used as validation.

At this time we are not making use of temporal coherence of tools. The predictions are based solely on the current frame. We haven't made use of any random distortions of input data such as flipping left to right, scaling, cropping or any other color distortions. During inference stage we can process, about 15 frames per second with NVIDIA GTX 650 Ti Boost graphics card.

References

[1] CATARACTS Challenge 2017. https://cataracts.grand-challenge.org/home/

[2] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Fei-Fei, Li, Imagenet: A Large-Scale Hierarchical Image Database, 2009 conference on Computer Vision and Pattern Recognition

[3] Rethinking the Inception Architecture for Computer Vision, http://arxiv.org/abs/1512.00567