CS230

Workflow Recognition in Cholecystectomy Videos

Suraj Menon, Ruben Mayer, Wei Cheung
šurajm72@stanford.edu, rmayer99@stanford.edu, cheungwz@stanford.edu

Abstract

Workflow recognition for surgery videos is an important piece of computer-assisted healthcare systems because it provides context about the state of the surgery. As a relatively new field, there is still many different learning frameworks that show promise for making significant gains in this computer vision challenge. For this project, we explored the effectiveness of Inflated 3D-CNNs, a recently developed video classification model, for workflow recognition on Cholecystectomy (gall-bladder surgery) videos. In addition, we built upon this method to develop a new method - Inflated 3D-CNNs + LSTM which adds extra temporal features to the I3D-CNN framework. While neither of these two methods outperformed the state-of-the-art for the Cholec80 data set, they did outperform some previously published research. This suggests that I3D-CNN is a promising new method that warrants more research.

1 Introduction

Workflow recognition in surgery videos is a relatively new area of study that has many practical uses in the healthcare space. Specifically, the ability to automatically detect the phase that a surgery is in from a video can help healthcare providers “monitor surgical processes, schedule surgeons and enhance coordination among surgical teams"[1]. In general, workflow recognition is essential for creating robust computer-assisted healthcare systems because they provide a key layer of context to the system.

For this project, we partnered with Stanford’s Technology Enabled Clinical Improvement Center (TECI) to build a workflow recognition model for surgical videos on Cholecystectomy surgeries. The goal was model that takes in a video input of a gallbladder surgery or heart surgery, and outputs time-labels that segment the video into the different surgery phases. Since workflow recognition is a relatively new field of research, there is no clear front-runner architecture for solving the task. Our contribution consisted of testing the effectiveness of a newer method for video classification, Inflated 3D-CNNs, based on the work by [6], as well as a new method that we developed, 3D-CNNs + LSTM, which involves adding extra temporal features at the end of 3D-CNNs network.

Our results showed that, while the Inflated 3D-CNNs model was unable to outperform the state-of-the-art methods developed by [1], it did outperform other published research for the same data set, such as the work by [3]. This suggests that, perhaps with more sophisticated optimizations, as well as more computational resources, the 3D-CNNs methodology might be useful for workflow recognition on surgical videos. Our own 3D-CNNs + LSTM framework did not overperform the "vanilla" 3D-CNNs model.

*

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Our work builds primarily off of SV-RCNet, developed by [1], which produced the state-of-the-art results for workflow recognition on the Cholec80 data set. In addition, we are building the methods developed by [6] to use Inflated 3D-CNNs to this problem.

SV-RCNet integrates a convolution neural network (CNN) and recurrent neural network (RNN) to capture the visual and temporal features from surgical videos. Specifically, a deep residual network (ResNet) and long short term memory (LSTM) were used to improve performance. SV-RCNet was trained on the Modeling and Monitoring of Computer Assisted Interventions (M2CAI) Workflow Challenge dataset and Cholec80 (cholecystectomy surgery) dataset. It outperformed the state-of-the-art method the time it was published. Our baseline is based on this LSTM + CNN approach.

In addition to [1]'s work, other significant research focused on the has been conducted for the M2CAI Workflow Challenge includes [2], [3], [4], [5]. [2] fine-tuned a ResNet with a temporal smoothing step that averages predictions temporally, and also used a Hidden Markov Model (HMM). [3] used a CNN architecture similar to the AlexNet, and employed transfer learning with features learned from ImageNet. The study also used contextual features and modeled the temporal occurrences of surgical phases by fitting Gaussian distributions. [4] used a four-stage approach including an AdaBoost classification and Hidden semi-Markov Model. [5] used CNN to extract visual features and experimented both the HMM and LSTM approaches to capture temporal features. We took reference in all these approaches, and decided to experiment on the CNN + RNN approach because it produced superior results.

[6] assessed multiple significant state-of-the-art approaches to general action recognition tasks on video data sets and developed a new approach by combining the key concepts in these approaches, and named it the Two-stream Inflated 3D-CNN. The approaches that were assessed in this paper include LSTM-CNN, 3D-CNN, Two-Stream (using RGB frames + optical flow frames) and 3D-Fused-Two-Stream. The new, combined model developed in this paper outperformed all these old approaches on the HMDB-51 and UCF-101 datasets, which are the two of the most cited datasets for action classification tasks. The paper also discussed a new Kinetic Human Action Video dataset (Kinetics) that the research team developed. The TS-I3D is currently widely considered as the state-of-the-art approach to action recognition. Since surgical workflow recognition is just one type of action recognition, we decided to experiment the I3D method on the Cholec80 dataset in this paper.

3 Dataset and Features

The dataset we are focusing on is the 'Cholec80' dataset. The dataset includes 80 videos of cholecystectomy operations (gallbladder extraction). Each of these operations varies in length, ranging between 20 minutes long to over an hour and a half. All the videos have 25 fps, and each frame is labeled into one of seven subclasses: Videos are at 25 fps. Each frame is labelled with one of seven subclasses: (1) preparation, (2) calot triangle dissection, (3) clipping and cutting, (4) gallbladder dissection, (5) gallbladder packaging, (6) cleaning and coagulation, and (7) gallbladder retraction. The videos have significant class imbalances, with some sub classes such as calot triangle dissection having many times more data available than other classes such as gallbladder dissection.

4 Methods

We implemented three different models to solve the workflow recognition task: A 2D CNN + LSTM baseline, an Inflated 3D-CNN model, and our own method, Inflated 3D-CNN + LSTM.

4.1 Baseline: 2D CNN + LSTM

Our LSTM + CNN implementation made use of the Keras framework and has three convolutional layers with a ReLU activation and max pooling, followed by a densely connected CNN layer with batch normalization and a ReLU activation function, followed by an LSTM layer with dropout and a tanh activation function, and two more Dense layers. Our final activation is a softmax function. We are using an adam optimizer with cross-entropy loss.

The largest challenge we faced when implementing this baseline involved memory management given our resource constraints. In order to train our model in a single GPU with 16GB of available memory, we experimented with several techniques that allowed us to satisfy these constraints while still making use of our entire data set. Specifically, we experimented with video channel down sampling, frame rate down sampling and changing clip size.

4.2 Inflated 3D-CNN

As discussed in Section 2, the I 3D CNN model developed by [6] currently has some of best performance results on general action recognition tasks. Our approach is to use transfer learning with the I3D model pre-trained on the ImageNet and Kinetics datasets, and train it on our Cholec80 data. From this pre-trained model we freeze several layers and train the final layers along with a new softmax classification layer with labelled data.

The model performs a series of 3D convolutions on a stream of RGB images and also upon a set of images distilled to their 'optical flow'. The optical flow images are preprocessed using a TV-L1 optical flow algorithm. The model itself uses 10 layers of an inception module containing two 3x3x3 layers and two 1x1x1 bottleneck layers with Max Pooling layers within and ending with an Average Pooling layer before the final prediction. The models original output is a softmax classification of actions from the UCF101 dataset. We processed the videos by cutting them up into 10s segments and then sampled at 5 fps. We then trained the RGB stream with croppings of the images at 224x224 and train the flow stream with images that have gone through the flow algorithm. We found that the running the flow algorithm through the entire dataset was very time consuming and so we have reported our results only from the RGB I3D pipeline.

To train the I3D RGB network we initially pre-processed the 60 videos in our train set into 10s clips spaced 30s apart. Each of these was saved individually and labelled with the phase of the surgery at the start of the clip. After initially generating the dataset we found though that there is significant class imbalance with class 1 (CalotTriangleDissection) and class 3 (Gallbladder Dissection) heavily biased. In order to counteract this, we added to the first dataset samples of the other phases sampled every 10s throughout the entire dataset and gave a much more even distribution.

Even so, there was still an imbalance upon a couple classes. In order to counteract this, we added a small modification to the usual softmax cross entropy loss function. To each loss calculated we would subtract a small proportional amount depending on the ratio the number of samples in a particular class to the total number of train samples. In order to even out the loss for the imbalanced classes, we then add a small boost to the loss equal to a heuristically determined value depending on the amount of the imbalance of the weakest class.

We then trained the network with a batch size of 32, using an ADAM optimizer with an alpha equal to .0001. To optimize the amount of training over the large data set, the training was conducted with only the last 'inception' layer of the network unfrozen. The rest of the variables retained the values from the pretrained imagenet training.

4.3 Inflated 3D CNN + LSTM

Though I3D in itself aims to learn temporal elements within 10s video clips, it does not have any sense of a larger picture. Due to the similarity of visual features within 10s clips spread throughout the surgeries, a further idea was to add an LSTM on top of I3D that would hopefully begin to recognize features that string together sequences of clips that more represent each phase. To try this, a simple LSTM cell with 16 hidden units was added at the end of the I3D model and was trained in an end-to-end manner with it.

5 Experiments and Results

From the 80 videos in the Cholec80 dataset, we assigned 60 videos in the training set, 10 videos in the development set and 10 videos in the test set. We ran all our experiments with the different models on the data according to the same train-dev-test split.

For our experiments we pre-processed the data in order to better fit the constraints of our computational systems. Specifically, we down sampled all the videos from 25 frames per second to 5 frames per

second. We then split each video into 10 second segments (with 50 frames each). In order to limit the amount of memory that we used, while still taking a varied sample of clips from the data set, we selected one 10 second clip out of every 30 seconds in the video. In addition, we down-sampled each clip from 3 channels to a single to grey scale channel for the CNN + LSTM portion of our project.

Method	Accuracy	F1-Score
2D CNN + LSTM	45.6%	39.8%
I 3D CNN	59.1%	54.2%
I 3D CNN + LSTM	34.5	18.1

Table 1: Performance of the three methods

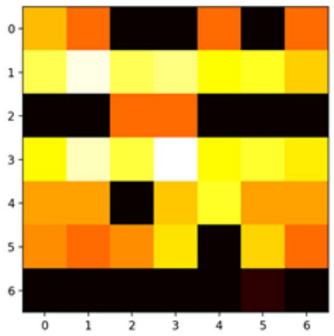

Figure 1: Confusion matrix for Inflated 3D CNN

Our results can be seen in Table 1 and the confusion matrix can be seen in Figure 1. Overall, the I 3D CNN method performed the best, with close to 60% accuracy in a 7-way classification task. While this is significantly below the 90% accuracy achieved by [1], it is on par with some research such as the work done by [3]. Our own method significantly under performed the baseline.

6 Analysis and Discussion

When compared to the state of the art, both our baseline and our implementation of the inflated 3D-CNN significantly under-perform the state-of-the art, such as [1] that achieves over 90% accuracy. This is to be expected, given that [1]'s work uses a very sophisticated set of modeling tools. [1] work is most similar to our baseline, however there are a few key differences. Specifically, [1]'s RCNet includes incorporates a deep ResNet model with over 50 layers, whereas we used a hand-built model with only 3 layers. In addition, [1] had the computational resources to make use of the entire data set, whereas in our work, we down sampled the frame rate and only used one 10-second clip out of every 30 seconds of video. Finally, [1] developed a novel prior knowledge inference methodology that allowed their models to take advantage of the natural structure of surgical videos.

Our 3D-CNN implementation also achieved lower performance than the state of the art, however at 60% accuracy, it is comparable to other research that tried novel video classification frameworks to classify this data set. For example, [3]'s work, mentioned earlier achieved only 52% accuracy.

Unfortunately, the 3D-CNN + LSTM method that we created was the worst performing. We believe that this is due in part to the class imbalance in the data set, given that the the model predicted all the data to a single category, which means that the LSTM features exacerbated this issue.

6.1 Error analysis

In performing analysis of the confusion matrix in the I3D results, there were some things to consider. This data set was particularly hard to work with from a purely visual analysis perspective due to the limited inter-phase variance and substantial intra-phase variance (as described in [1]). We looked at

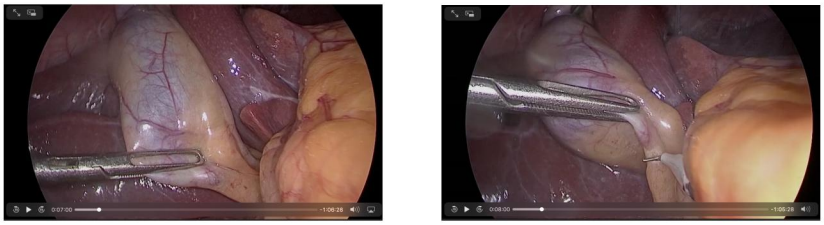

Figure 2: Left: 'Preparation' phase mislabelled as 'CalotTriangleDissection'— Right: 'Correct labelling of 'CalotTriangleDissection'

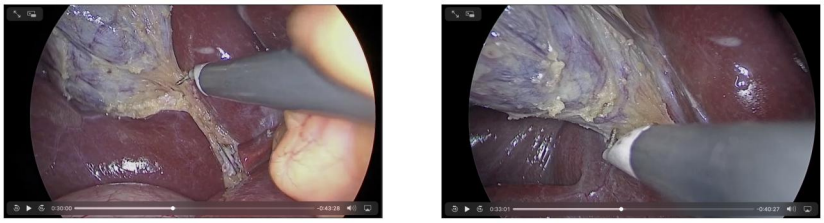

Figure 3: Images with similar visual features confuse the network due to lack of inter-phase difference

specific video frames that were misclassified and as we expected found there to be the aforementioned variance problems.

In these two images, we see how there is very little variation from Phase 0 (Preparation) to Phase 1 (CalotTriangleDissection). Due to the natural imbalance of samples from Phase 0, we see that these type of Phase 0 samples are often misclassified as Phase 1.

In these two images, we see how two images in Phase 3 (Gallbladder Dissection) are classified as Phase 1 (CalotTriangleDissection) and Phase 3 respectively. The images have very similar visual features, but it is likely that since the same tool is used in Phase 1 and in Phase 3 the network has difficulty differentiating between the two.

7 Conclusion and Future Work

In conclusion, from these results we believe that the I3D network is a viable solution for workflow analysis for Cholecystectomy surgeries and potentially others. The accuracy achieved shows promise that if trained with more unfrozen layers and for a longer period of time on the more class balanced dataset that I3D could serve as a good base for the visual analysis of this dataset. We also would like to finish processing the dataset and correctly train the 'flow' pipeline and see if that improves our results when combined with the I3D RGB results.

Apart from improving the I3D implementation, other work we would like to continue is to find a better method of using a sequence model to tie features together one from 10s clip to another. We are hoping that a technique like this may be able to better capture the transitions from phase to phase.

8 Contributions

Suraj worked on building the i3d model, doing error analysis, experimenting with different pre-processing hyper-parameters (fps) and different class balance. Ruben worked on building the baseline models with 2D CNN and LSTM, and also the first data pre-processing work. Wei worked on the building the Optical Flow stream of the i3d model and initial project scoping.

For our work, we built upon open source code as well as support from our mentor for our implementation of our models. This code is available in the following links:

Our code: https://github.com/surajm72/cs230-surgical-group

Referenced code: https://github.com/deepmind/kinetics-i3d https://github.com/LossNAN/I3D-Tensorflow https://github.com/inubushi/LSTM-KTH

References

[1] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. Fu, P. Heng. 2018. *SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network* IEEE Transactions on Medical Imaging, Vol 37, No. 5. Pages 1114-1126.

[2] R. Cadène, T. Robert, N. Thome, and M. Cord. 2016. *M2CAI workflow challenge: Convolutional neural networks with time smoothing and hidden Markov model for video frames classification.* https://arxiv.org/abs/1610.05541

[3] M. Sahu, A. Mukhopadhyay, A. Szengel, S. Zachow. 2016. *Tool and Phase recognition using contextual CNN features* arXiv:1610.08854v1.

[4] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin. 2016. *Data-driven surgical workflow detection: Technical report for M2CAI 2016 surgical workflow challenge* http://camma.u-strasbg.fr/m2cai2016/reports/Dergachyova-Workflow.pdf

[5] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy. 2016. *Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016.* https://arxiv.org/abs/1610.08844

[6] J. Carreira, A. Zisserman. 2018. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset* arXiv:1705.07750v3.

[7] K. Simonyan, A. Zisserman. 2014. *Two-Stream Convolutional Networks for Action Recognition in Videos*

[8] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, F. Li. 2018. *Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks* arXiv:1802.08774v2.