Journal of Big Data

RESEARCH Open Access

Towards more efficient CNN-based surgical tools classification using transfer learning

Jaafar Jaafari[1*], Samira Douzi[2*], Khadija Douzi[1*] and Badr Hssina[1*]

*Correspondence: jaafar.jaafari@etu.fstm.ac.ma; s.douzi@um5r.ac.ma; khadija.douzi@fstm.ac.ma; badr.hssina@fstm.ac.ma
[1] FSTM, University Hassan II, Casablanca, Morocco
[2] FMPR, University Mohammed V, Rabat, Morocco

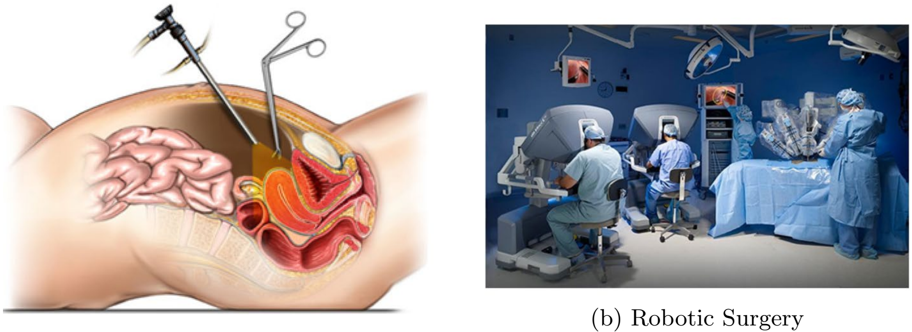
Abstract

Context-aware system (CAS) is a system that can understand the context of a given situation and either share this context with other systems for their response or respond by itself. In surgery, these systems are intended to assist surgeons enhance the scheduling productivity of operating rooms (OR) and surgical teams, and promote a comprehensive perception and consciousness of the OR. Furthermore, the automated surgical tool classification in medical images is a real-time computerized assistance to the surgeons in conducting different operations. Moreover, deep learning has embroiled in every facet of life due to the availability of large datasets and the emergence of convolutional neural networks (CNN) that have paved the way for the development of different image related processes. The aim of this paper is to resolve the problem of unbalanced data in the publicly available Cholec80 laparoscopy video dataset, using multiple data augmentation techniques. Furthermore, we implement a fine-tuned CNN to tackle the automatic tool detection during a surgery, with prospective use in the teaching field, evaluating surgeons, and surgical quality assessment (SQA). The proposed method is evaluated on a dataset of 80 cholecystectomy videos (Cholec80 dataset). A mean average precision of 93.75% demonstrates the effectiveness of the proposed method, outperforming the other models significantly.

Keywords: Minimally-invasive-surgery, Deep Learning, Computer-vision, Transfer learning, Data augmentation

Introduction

Minimally invasive surgery (MIS) techniques had a massive effect on operations from the 1990s. This progression in operative techniques improved the satisfaction of patients enduring surgical interventions. MIS surgeons operate with minimal damage to the body than with open surgery. Minimally invasive surgery (Fig. 1a) is now the common surgical method in plentiful interventions, like cholecystectomy(removal of the gallbladder). It is effectuated by three to four holes(a few millimeters long) in the abdomen, instead of a large overture in the open surgery. The laparoscope is inserted in one incision, while the surgical instruments are entered into the others. The laparoscope is the main instrument in the minimally invasive surgery, which is a thin tool with a tiny high-resolution camera at the end. The image stream of the high-resolution camera is broadcast on a screen in the operations room.



(a) Laparoscopic surgery

(b) Robotic Surgery

Fig. 1 Minimaly invasive surgery type

Minimally invasive surgery has myriad advantages than conventional open surgery [1, 2]:

- *Smaller incisions*: The hysterectomy is an operation to remove the uterus. It requires between 152mm and 300mm incision using open surgery. On the other hand, using minimally invasive surgery requires only 5 to 15 millimeters only.
- *Less pain*: Trauma caused by cutting large sections of muscle and tissues in open surgery is unnecessary. MIS is characterized by its small surgical tools, requiring smaller scars and causing less pain and faster healing.
- *Reduced Risk of complications*: The risk of blood loss, which can take the patient's life, is reduced when using smaller incisions [3].
- *Shorter hospital stays*: Open surgery patients need 2 to 3 nights stay, while MIS patients can go home after only one night at the hospital.
- *The visual results*: MIS is done with very little incision. Scars are undetectable after a couple of months.

On the other hand, learning MIS procedures have multiple limitations, such as:

- The operative area is displayed on the monitor as a 2-D image, instead of the regular 3-D eyesight. This absence of depth perception is a challenge for surgeons.
- Limited training chances because of patient safety and resource concerns.

To overcome these issues, the surgical video stream is recorded and exploited. These videos are utilized in retrospective analysis and postoperative Surgical Quality Assessment (SQA), which are investigating the recorded videos narrowly, to detect possible mistakes and evaluate the surgeon expertise. They are also used as academic material for debutante surgeons [4]. Furthermore, recording the surgical intervention is compulsory in many countries [5], and provided as proof in divers circumstances.

However, MIS videos tend to last several hours. Thus, navigation and searching through these videos are cumbersome and time-effort consuming. To overcome this problem, we propose a deep learning-based solution to classify surgical tools in MIS videos and stock the results in a database, to execute specific queries, and allow amateur surgeons and postoperative controllers, to navigate and access to desired video segments easily.

In this study, we designed a system for the classification of surgical tools in MIS videos utilizing advanced deep neural networks. The system consists of the following steps:

Firstly, preprocessing the digital images which includes splitting the Cholec80 videos in one frame per second images, then resizing these images to a fixed size.

The next essential step includes an augmentation process to overcome the unbalanced data problem by increasing the size of minority classes.

Finally, we use a well-known deep learning model to train it on the augmented data images. Then, we compare the proposed approach with other methods.

This paper is organized as follows: Sect. "Related works" presents the related works, Sect. "Deep learning and computer vision" defines the basic concepts of the deep learning and computer vision, Sect. "Our approach" describes the detailed methodology of the proposed approach, including the implementation and experimental results, whereas sect. "Conclusions and future works" covers the conclusion of the paper.

Related works

Classification, segmentation, and tracking using convolutional neural networks (CNN) have made state-of-the-art results in the field of medicine, for example, pulmonary tuberculosis segmentation [6], brain tumor segmentation [7], stroke lesion segmentation [8], breast cancer classification [9], and artery and vein classification [10].

The purpose of cholecystectomy is to remove the gallbladder: this operation can be performed laparoscopically and monitored through an endoscope. The recorded videos were used for multiple purposes. We summarize some uses of computer vision in minimally invasive surgery videos in:

- *Laparoscopy*: Kletz [11] applied a region-based convolutional neural network (R-CNN), to recognize the surgical instruments in laparoscopy. A custom dataset generated from laparoscopic gynecological videos was used. Amy [12] combined a region-based convolutional neural network (faster R-CNN) and VGG16 to detect laparoscopic surgical tools and perform an operative skill assessment. The dataset used is M2CAI. Christian [13] performed the segmentation of surgical instruments in laparoscopic surgery using a self-supervised method based on the kinematic model of the robot as a source of information. A fully convolutional neural network (FCN) was used on VIVO dataset. This dataset was obtained from a robotized endoscopy system. Choi [14] choose a convolutional neural Network called YOLO (You Only Look Once) to perform the surgical tools detection in laparoscopic surgery on the M2CAI 2016 challenge dataset. Wang [15] proposed an ensemble learning approach based on VGGNet and GoogleNet. The multi-label classification of the surgical instruments was tested on M2CAI dataset. Atttia.m [16] used a hybrid CNN-RNN AutoEncoder-Decoder to segment surgical tools. The approach was tested on MICCAI 2016 endoscopic vision challenge dataset.
- *Robotic surgery* (Fig. 1b: Shvets [17] presented a deep learning-based solution for robotic instrument segmentation. The dataset used is MICCAI 2017 endoscopic vision sub challenge [18]. Islam [19] proposed a real-time instrument segmentation

tool in robot-assisted minimally invasive surgery (RMIS) using Multiresolution Feature Fusion (MFF) block and a light-weight CNN to identify the surgical tool. This approach was tested on MICCAI 2017 dataset. Shvets [20] presented an instrument segmentation framework, in robotic surgery, using adversarial learning, Multi-resolution Feature Fusion (MFF), and a fully convolutional network (FCN). Colleoni [21] applied spatiotemporal layers using a fully convolutional neural network (FCNN), to perform robotic surgical tool detection and articulation estimation. EndoVis challenge 2015 dataset was used. Automatic instrument segmentation in robot-assisted surgery was implemented by Shvets [20] using U-Net, TernausNet, LinkNet, VGG11, and VGG16 encoders, on a custom dataset obtained from DA Vinci Xi surgical system. Sarikaya [22] performed the detection and localization of robotic tools in robot-assisted surgery videos using region proposal network (RPN), on ATLAS Dione dataset.

- *Other uses of MIS videos*: As we mentioned earlier, minimally invasive surgery videos often have a duration of several hours. That's why Chittajallu [23] presented a content-based video retrieval system, similar to a query image. The author used a CNN model called ResNet50, trained with a siamese triplet, ranked and refined with IQR (iterative query refinement), based on the user feedback. The Model can detect the surgical instrument in the query image and search for similar frames in the stocked MIS videos.

The preoperative surgery duration prediction is done manually, thus, the surgeons underestimate surgery duration by 31 min [24] on average, causing a longer waiting time for patients and a non-optimized exploitation of the operation room. Twinanda [25] proposed an automatic method, using the visual information from laparoscopic video, to detect instruments corresponding to a specific phase, and it continuously predicts remaining surgery duration, without any human intervention. A CNN is used to extract visual features from the video frames, and LSTM (long-short term memory) was used to predict the remaining surgery time. The results were very positive, with 15.2 ± 4.7 minutes for short videos, 12.5 ± 4.6 minutes in medium videos, and 23.1 ± 9.4 minutes in long videos.

Deep learning and computer vision

Machine learning (ML), unlike other types of computer programming, provides the ability to automatically learn and improve from experience without being explicitly programmed. ML focuses on the development of computer programs that can gather data and use it to learn. Deep Learning (DL) is a member of the ML family and has seen a qualitative leap in the past years, driven by the availability of large datasets and the evolution of computer resources. This field has witnessed imminent progress in the ability of machines to understand and manipulate data, including videos and images. DL is based on artificial neural networks, that imitate the human brain in the decision-making process by generating patterns.

Some of the biggest hits of deep learning have been in the area of computer vision (CV). CV is a field of artificial intelligence that trains computers to interpret and understand the visual world from digital images or videos. Computer vision is also defined as an area of study that seeks to develop techniques to help computers "see" and understand the content of digital images. When the computer receives an image as an input, it's an array of pixel values (Fig. 2), each of these values is a number between zero and 255, which tells the intensity of the pixel, meaningless to humans, but it is the only input to machines.

Deep Learning achieved revolutionary results on challenging computer vision problems such as image classification [26], object detection [27] and face recognition [28]. Medical imaging is one of the most benefiting fields from the evolution of computer vision, it helps doctors by alerting them when an area in an image is doubtable (e.g. detecting a tumor in an image).

Convolutional neural networks (CNNs) [29] is a specific class of neural network (NN) that was created to learn visual features from images. Nowadays, it is the most successful deep learning approach to handle the image classification task [30].

The standard CNN is basically composed of three layers: convolutional, pooling, and fully connected layers.

- Convolutional layer: is the core building block of a CNN, it does massive math operations. The picture is convoluted with a specific filter to extract the desired features. It is the first action is a CNN, the result is an activation map or a feature map.
- Pooling Layer: aims to reduce the dimensions of the feature maps. For each feature map received from the previous layer, a filter is applied to summarize the features lying within the region covered by the filter.
- Fully connected layer: is the last layer, and the classification result is specified based on the category (or categories) with highest value. At this stage, the output of the previous layer is flattened and turned into a single vector, then assigning weights to predict the correct label and giving the final probabilities for each category.

In the medical field, CNN is employed quite successfully in detecting and tracking surgical tools (Fig. 3). CNN obtain an image as input and transform it to a vector and apply some simple operations like convolution, pooling, then, output a probability of potential image classes.

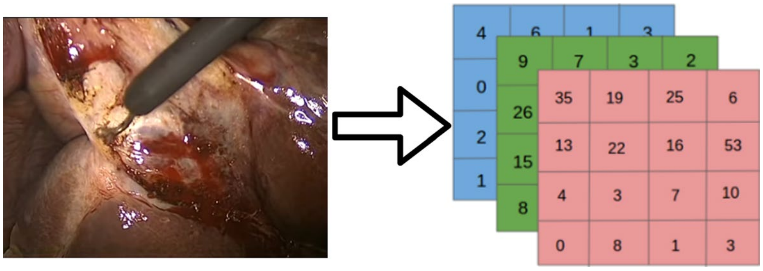


Fig. 2 Image to an array of pixels values

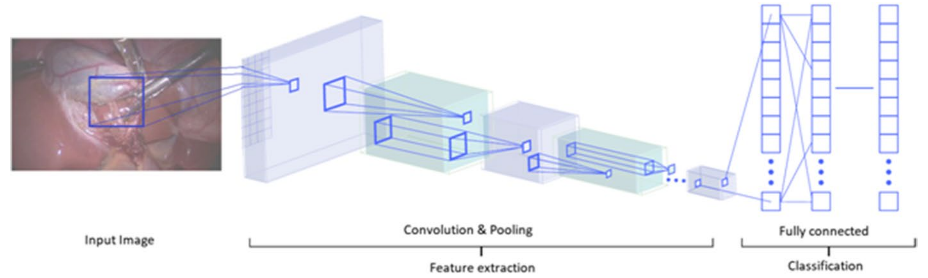


Fig. 3 Convolutional Neural Network architecture

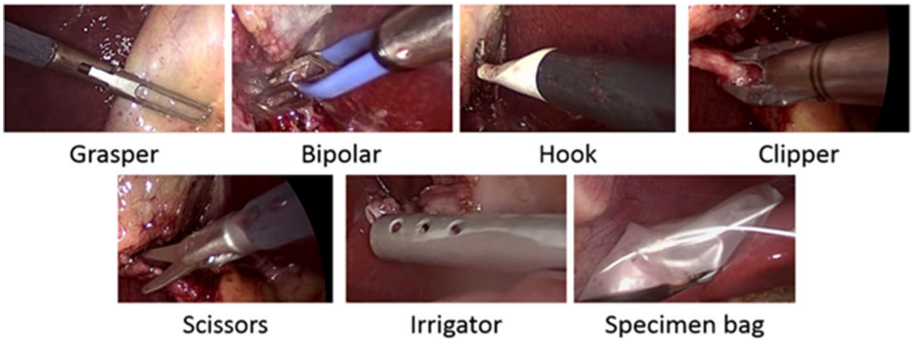


Fig. 4 Seven surgical tools included in the CHOLEC80 dataset

Our approach

This section includes the dataset description, data preprocessing, data augmentation, network architecture of the proposed model, the experimental results, and the discussion.

Data preprocessing

Dataset

Cholec80 [31] is a cholecystectomy surgery videos dataset containing 80 videos, performed by 13 surgeons. Video resolution is 1920 × 1080 pixels with 25 Frames Per Second(fps) as frame rate. Video length is varied between 12 min 19 s (minimum) and 1 hour 39 min 55 s (maximum), with 38 min and 26 s on average and more than 51 hours of surgery in total. Cholec80 is fully annotated with image-level surgical tool labels for binary detection.

In Cholec80, seven tools were used and annotated (Fig. 4) show an example of the seven surgical tools present in the dataset, namely: specimen bag, bipolar, scissors, clipper, hook, grasper, and irrigator. As the images are collected using different laparoscopes and from different surgeons, they come with different angles and resolutions, and sometimes they have a poor resolution, focus, or blurred. In addition, the tool is labeled as present if half of it appears in the image. One binary label is provided per image and per tool as an annotation (Multilabel classification).

Data preprocessing

Videos are processed using FFmpeg 3.0 and all video streams are encoded with libx264, using 25 frame per second (FPS).

Firstly, the video width is scaled to 480, and the height is determined to maintain the aspect ratio of the original input video. Next, the audio is isolated from all videos.

Since videos are brut and not edited, they have several empty and irrelevant frames at multiple scenes (beginning and the end of the videos). Furthermore, these frames are noisy and computationally expensive. Therefore, we cut these nonrelevant frames using a background detection model. The latter was trained to identify unimportant segments that were captured outside the body. Next, these frames are used to recognize the real start and end of the surgery in the original video and cut it down.

This step and the final verified video files are automatically processed and stored in local computer.

Since the Cholec80 dataset is labelled with 1 fps tool presence annotation, we split the preprocessed videos in 1 frame per second images. Splitting videos is a display technique that reposes on fractioning the video into images.

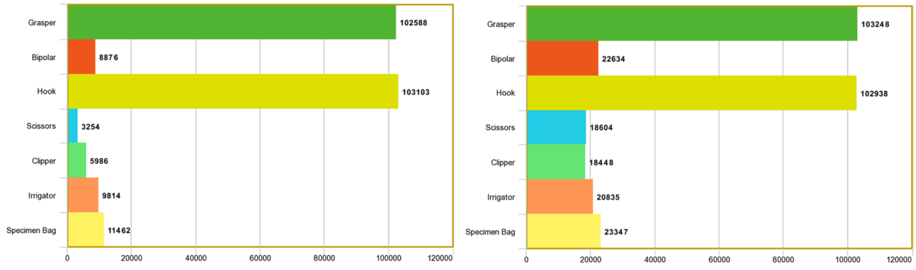
Finally, and given that neural networks receive inputs of the same size, all images need to be resized to a fixed size before inputting them to the CNN. Moreover, the image dimension is often reduced in order to fit a reasonably sized batch in GPU memory. Thus, each image is resized to 250 × 250pixels.

Data augmentation

In Cholecystectomy surgery, some tools are used more frequently than others. Consequently, Cholec80 video frames belonging to those tools outnumber the video frames belonging to the other tools, leading to unbalanced data. This issue affects the generalization of the model and reduces the CNN efficiency to classify the different tools.

To overcome this problem, image augmentation techniques are used to increase the size of the minority classes. Images are augmented by affine transformations and blurring (Fig. 5). We consider those transformations that preserve tool presence, like:

- Rotation: Minority class images are rotated at an angle of zero, 40, 85, 125, 250, and 300.
- Mirroring: Mirror the image along the x-axis and y-axis.
- Shearing: The images were shifted at 40 degrees in the counter-clockwise direction.



BEFORE AUGMENTATION

AFTER AUGMENTATION

Fig. 5 Tools presence before and after augmentation

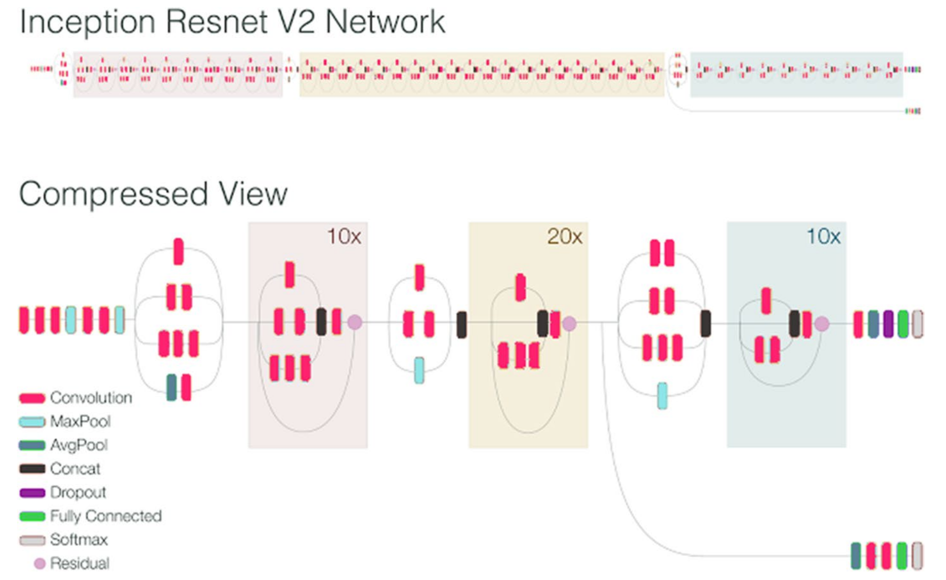


Fig. 6 InceptionResnet-v2 architecture

- Padding: Padding 5px on each border, using the reflect mode, which pad with the reflection of image without repeating the last value on the edge.

As shown in Fig. 5, the total number of images before the this phase was 245086 frames, and after the augmentation it becomes 310054 frames.

Network architecture

InceptionResnetV2

The residual neural network (ResNet) is an artificial neural network that was probably the most revolutionary work in computer vision/deep learning in recent years. Resnet made a state-of-art results on many computer vision applications, such as object detection and face recognition. From Alexnet [32] in 2012, that win the SVRC2012 classification contest, Researchers focused on developing Deep Residual Networks. VGG Network (19 layers) and GoogleNet (22 layers) are state-of-arts CNN architectures. They tried to go deeper and deeper, however, piling layers does not mean increasing network depth. The vanishing gradient problem makes the neural network very hard to train, that's why Resnet introduce "Identity Shortcut Connection" , to skip layers. The authors of [33] argue that stacking layers shouldn't degrade the network performance, because we could simply stack identity mappings (layer that doesn't do anything) upon the current network, and the resulting architecture would perform the same.

Another neural network that was a milestone in the CNN development is the Inception network. CNNs tends to stack convolution layers. Unlike CNNs, inception uses other solutions to get a better performance in terms of speed and accuracy.

InceptionResnet-V2 (Fig. 6) is a hybrid inception network that combines inception and residual networks (both are SOTA architectures), to boost the performance.

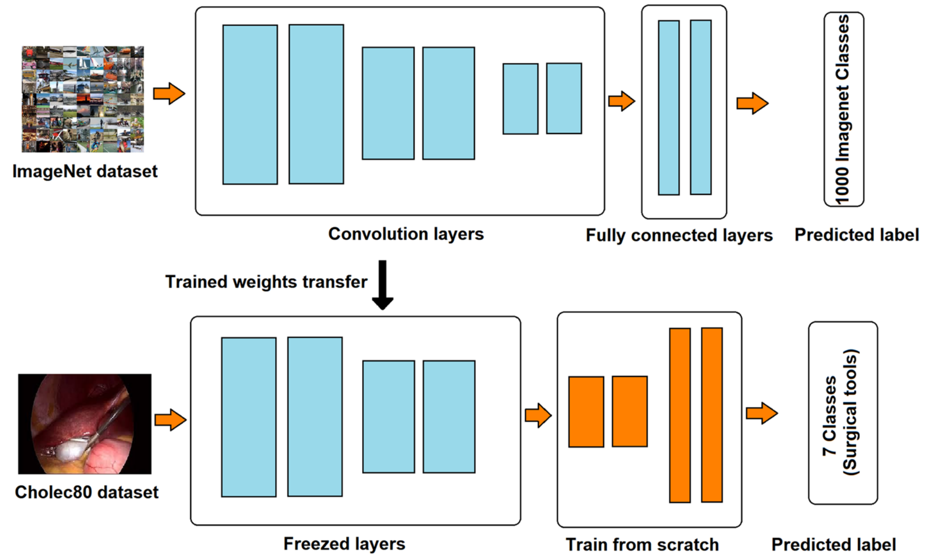


Fig. 7 Illustration of fine-tuning a convolutional neural network pretrained on ImageNet dataset

InceptionResnetV2 is trained on more than one million images from the ImageNet dataset [34]. The network has a default input size of 299-by-299.

Transfer learning

Transfer learning is taking a network pretrained on a dataset and apply it to recognize new image/object categories. Essentially, we can exploit the robust, discriminative filters learned by state-of-the-art networks on challenging datasets (such as ImageNet or COCO), and use these networks to recognize objects the model was never trained on.

In deep learning, feature extraction and fine-tuning are two types of transfer learning:

- Transfer learning via feature extraction is done by freezing all the convolutional neural network layers, and changing only the classification layer, which is the final layer. The pretrained network is used to extract the input image features. This technique is used when the new data are similar to the original training dataset (Imagenet).
- On the other hand, fine-tuning requires more modifications than feature extraction (Fig. 7). The layers are initialized with the pre-trained neural network model weights. The model architecture is updated by removing the fully connected layer heads and replacing it with a new one, then training it to predict the input classes. Furthermore, some of the last layers could be unfrozen, in order to perform a second pass of training. Freezing means that these layer weights will not be updated in the training process. This technique is used when data similarity is low between the original training dataset (Imagenet) and our images. That is why, in our case we fine-tuned Inception-ResnetV2.

Fine-tuning inceptionResnetV2

In the fine-tuning approach, the representations learnt by the previous network are used to extract the meaningful features of the new dataset images and the activation maps generated from the last convolutional layer are fed to the newly constructed fully connected network which acts as the classifier.

Therefore, the first step is to truncate the fully connected node at the end of the pretrained network (Softmax layer), and change it with a new freshly initialized sigmoid layer that is compatible with our multilabel classification task. This will predict a probability of class membership for the seven labels and assign a value between 0 and 1. The sigmoid function is calculated as :

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x) \tag{1}$$

After leaving off the fully connected (FC) head of InceptionResnetV2 that contains 1000 classes (the 1000 output classes of the ImageNet dataset), we construct a new FC layer (classifier layer), with seven classes (surgical tools number), and append it to our model.

Next, in order to learn very generic features, we freeze the first early blocks. Moreover, it lets the network capture common features like edges and curves that are applicable to our new classification task. Additionally, this step ensures that any previous robust features learned by the CNN are not destroyed.

Then, we train the new FC head that is connected to the model to take the lower level features from the front of the network and map them to the desired output classes. Once this has been done, we unfreeze some layers of the top layer of the frozen model, by setting these layers as "trainable=True", and continue training, so that in further SGD epochs their weights can be fine-tuned for the new task too. Furthermore, we used a smaller learning rate to train the network because we expect that the pretrained weights are quite good already as compared to randomly initialized weights.

Finally, the input size of the model was changed to (250, 250, 3), with 250 * 250 as our frame dimension and 3 is our color channels (RGB).

Experimental results and discussion

Implementation parameters

As we mention in section A-1, we used Cholec80 for performance evaluation. 60 videos were assigned to the training set (producing 241 842 image), while 20 videos (producing 68 212 images) were assigned to the test set. The grasper and hook appear more often than other tools, leading to unbalanced data. Image augmentation techniques are applied to overcome this problem, as described in the A-2 section.

We fine-tuned InceptionResnet-v2. The latter is pretrained on ImageNet dataset. The fine-tuning process is defined in an earlier section.

The tool detection is multilabel and multiclass task, as different tools can be present at the same time. Our model is trained using stochastic gradient descent, while binary Cross-Entropy is used as a loss function and Sigmoid is used as a final activation function.

The training process was run for 70K iterations on a batch size of 32 images, with an initial learning rate of 0.001, that we decay it by a factor of 10 after 12K iterations.

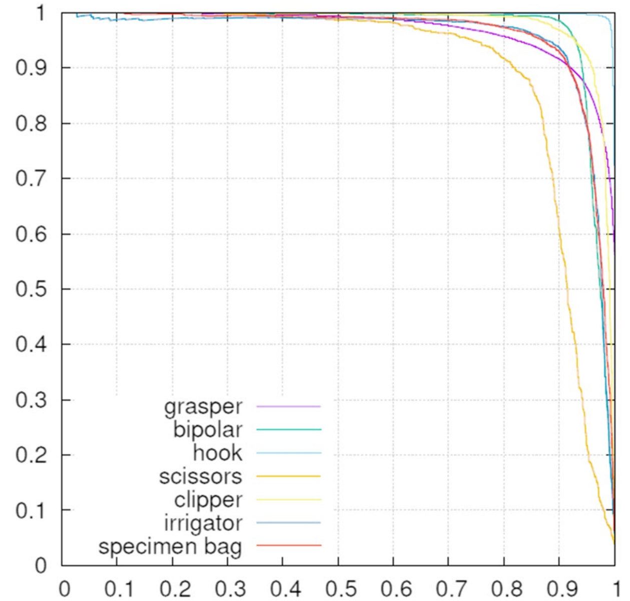


Fig. 8 Precision-recall (PR) curve

The network is trained on Intel®Core^{TM} i7-9700K processor, 16GB memory and NVIDIA Geforce GTX 2080.

Results and discussion

We implemented Convolutional Neural Network for surgical tool classification. This network is trained on Cholec80, which is a cholecystectomy surgery videos dataset containing 80 videos.

Figure 5 shows that the instances of surgical tool presence in laparoscopy is massively unbalanced. Among the seven tools, the grasper and the hook are present majority of the time, although scissors has very fewer samples. Giving any network such unbalanced data polarizes the network to the surgical tools which have high number of samples. Therefore, we used some augmentation techniques in order to overcome this problem.

The performance of our model is measured by the average precision (AP) metric, presented in Eq. (2). It is a popular metric in measuring the accuracy of object detectors. Each frame is annotated with a single or multiple surgical tools. Accuracy is calculated by comparing the ground truth annotations with the predicted labels. Additionally, we used Precision-Recall curves. Precision-Recall curves are used instead of the ROC(Receiver Operating Characteristic) curves when there is a class unbalance (Fig.8).

The average precision (AP) is calculated as:

$$AP = \frac{1}{k} \sum_{Recall_i} Precision(Recall_i) \tag{2}$$

where

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Table 1 Average precision (AP) for all tools, evaluated on Cholec80 dataset

Surgical tool	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	SpecimenBag	Average
Average precision	96.58	95.04	99.68	81.24	95.11	93.71	94.92	93.75

Table 2 Comparison of the Average precision (AP) for all tools with other models

Tool	M.Sahu [35]	EndoNet [31]	Amy.J [12]	Jo [36]	Kanakatte [37]	Our Model
Grasper	73.9	84.8	87.2	92.1	93.8	96.58
Bipolar	40.8	86.9	75.1	82.3	90.0	95.04
Hook	95.1	95.6	95.3	85.9	86.1	99.68
Scissors	26.2	58.6	70.8	81.2	100	81.24
Clipper	35.3	80.1	88.4	85.3	91.9	95.11
Irrigator	33.2	74.4	73.5	82.9	88.4	93.71
SpecimenBag	76.6	86.8	82.1	83.2	83.1	94.92
Average (mAP)	54.44	81.0	81.8	84.7	90.5	93.75

and

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

and k is the number of points interpolated in the Precision-Recall curve,

TP = True Positives, FP = False Positives, and FN = False Negatives.

The performance of each model is calculated by the mean Average Precision (mAP), which is the mean of the average precision score for each surgical tool. It is calculated as :

$$mAP = \frac{1}{7}\sum_{i=1}^{7} AP_i \tag{5}$$

where 7 is the number of surgical instruments.

The classification performance for all instruments using the augmented data is shown in Table 1. It specifies the average precision of the trained model for classifying the seven surgical tools.

In this work, we overcome the unbalanced data problem of the Cholec80 dataset. However, training the model to recognize minor classes (e.g., scissors, irrigator) was a real challenge. It can be seen that the hook obtains the greatest average precision. Two possible explanations are that it has the highest samples number, and it has a unique shape, making it easily recognizable from other tools. Moreover, the specimen bag, bipolar, clipper, and the grasper performed well. Irrigator is often misclassified, maybe due to its universal shape along with infrequent and irregular presence. The latter is an instrument used for flushing and suctioning a space, it is typically used essentially when blood concentrates in the surgical area. When taking a closer look at the average precision between these instruments and scissors, remarkable differences can be regarded, reflecting the fact that the latter has the lowest sample number and its common two-pronged shape.

To evaluate the performance of the proposed model, we compared the results with those of previous studies. Table 2 outlines and compares our model with the best performing related models. To differentiate our model from the models, we used the average precision computed for all classes, and the mean average precision for each model.

Sahu [35] and Twinanda [31] fine-tuned AlexNet on Cholec80 dataset, without any data augmentation techniques. While Amy [12] used VGG16, which is is an extension of Alexnet, containing eight weight layers, with a new architecture including 16 weight layers, and the authors performed data augmentation by randomly flipping frames horizontally. On the other hand, Jo [36] used YOLO9000 with motion vector prediction, and Kanakatte [37] approach was based on Resnet.

It can be seen that our model yields significantly better results than Sahu [35], EndoNet [31], Amy [12], and Jo [36] architectures, especially in the scissors, which is the less represented class in Cholec80. This may be due to the multiple augmentation techniques used. We can also observe that our model outperforms all models in classifying almost all the surgical tools. Two potential justification for this: in the pre-processing phase, we used more augmentation techniques, that's justify what the average precision is higher. In addition, InceptionResnet-v2 is network of 164 layers deep, having much more convolutional layers than VGG16 and AlexNet. Each convolutional layer's parameters consist of a set of learnable filters, allowing the network to learn much more information.

The results demonstrated that leaving off the classic pre-processing techniques improved the classification outcomes. This is principally crucial in the case of classification projects with highly unbalanced data. Data augmentation overcomes this issue, leading to a higher average precision than the models trained without data augmentation. However, one drawback of adding more data is the demand on resources and computational intensiveness, which can increase the time of total data preparation and training time.

Conclusions and future works

In this research, we overcome the unbalanced data problem in the publicly available laparoscopy video dataset Cholec80. We proposed multiple data augmentation techniques. Moreover, we exploited the techniques of transfer learning, especially the fine-tuning approach. Then, we choose to train the augmented data with a fine-tuned Inception-Resnet-v2 network. The latter relies on convolutional neural networks (CNNs) and it is pretrained on ImageNet dataset. Our Model has shown that the classification task works reliably. It was observed that the proposed model outperform the other methods in terms of classifying the surgical tools with an accuracy of 96.58, 95.04, 99.68, 81.24, 95.11, 93.71, and 94.92% for the surgical instruments Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator and SpecimenBag, and an average mean precision of 93.75%. Due to this high average precision of our model, it can be used for Computer-assisted instruction (CAI) as a basis of automatic surgical video indexing.

In future works, we will test some other neural network architectures, such as NASNet, DenseNet, and EfficientNet, and try it on other surgical datasets, with other augmentation data techniques. Furthermore, we plan to investigate the impact of the augmentation data phase on the processing time.

Acknowledgements
Not applicable.

Authors'contributions
All authors read and approved the final manuscript.

Funding
Not applicable. This research received no specific grant from any funding agency.

Availability of data and materials
Cholec80 dataset: http://camma.u-strasbg.fr/datasets. For any collaboration, please contact the authors.

Declarations

Ethics approval and consent to participate
The author confirms the sole responsibility for this manuscript. The author read and approved the final manuscript.

Consent for publication

Competing interests
The authors declare that they have no competing interests.

Received: 20 May 2021 Accepted: 20 August 2021
Published online: 30 August 2021

References

1. Tim Xu, Hutfless Susan M, Cooper Michol A, et al. Hospital cost implications of increased use of minimally invasive surgery. JAMA Surg. 2015;150(5):489.
2. Chen Q, Merath K, Bagante F, Akgul O, Dillhoff M, Cloyd J, Pawlik TM. A comparison of open and minimally invasive surgery for hepatic and pancreatic resections among the medicare population. J Gastrointest Surg. 2018. https://doi.org/10.1007/s11605-018-3883-x.
3. Ee WWG, Lau WLJ, Yeo W, Bing VY, Yue WM. Does minimally invasive surgery have a lower risk of surgical site infections compared with open spinal surgery? Clinical. 2013.
4. Mota P, Carvalho N, Carvalho-Dias E, Joãao Costa M, Correia-Pinto J, Lima E. Video-based surgical learning: improving trainee education and preparation for surgery. J Surg Edu. 2018;75(3):828–35. https://doi.org/10.1016/j.jsurg.2017.09.027.
5. Henken KR, Jansen FW, Klein J, Stassen LPS, Dankelman J, van den Dobbelsteen JJ. Implications of the law on video recording in clinical practice. Surg Endosc. 2012;26:2909–16. https://doi.org/10.1007/s00464-012-2284-6.
6. Li L, Huang H, Jin X. AE-CNN Classification of Pulmonary Tuberculosis Based on CT Images. 2018 9th International Conference on Information Technology in Medicine and Education (ITME); 2018. https://doi.org/10.1109/itme.2018.00020.
7. Xiao Z, Huang R, Ding Y, Lan T, Dong F, Qin Z, Wang W. A deep learning-based segmentation method for brain tumor in MR images. 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS); 2016. https://doi.org/10.1109/iccabs.2016.7802771.
8. Joshi S, Gore S. Ishemic Stroke Lesion Segmentation by Analyzing MRI Images Using Dilated and Transposed Convolutions in Convolutional Neural Networks. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA); 2018. https://doi.org/10.1109/iccubea.2018.8697545.
9. Ye J, Luo Y, Zhu C, Liu F, Zhang Y. Breast cancer image classification on WSI with spatial correlations. ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. https://doi.org/10.1109/icassp.2019.8682560.
10. Kiruthika M, Swapna TR, Kumar SC, Peeyush KP. Artery and Vein classification for hypertensive retinopathy 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).
11. Kletz S, Schoeffmann K, Benois-Pineau J, Husslein H. Identifying Surgical Instruments in aparoscopy using deep learning instance segmentation. 2019 International Conference on Content-Based Multimedia Indexing (CBMI); 2019. https://doi.org/10.1109/cbmi.2019.8877379.
12. Jin A et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE; 2018
13. da Costa Rocha C, Padoy N, Benoit R. International Conference on Robotics andAutomation (ICRA) Palais des congres de Montreal, Montreal, Canada, 20–24. Self-SupervisedSurgical Tool Segmentation using Kinematic Information; 2019.

14. Choi B, Jo K, Choi S, Choi J. Surgical-tools detection based on Convolutional NeuralNetwork in laparoscopic robot-assisted surgery. 2017 39th Annual International Conference of theIEEE Engineering in Medicine and Biology Society (EMBC); 2017.
15. Wang S, Raju A, Huang J. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos; 2017.
16. Attia M, Hossny M, Nahavandi S, Asadi H. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017.
17. Automatic instrument segmentation in robot-assisted surgery using deep learning,"A. A. Shvets. 2018 17th IEEE International Conference on machine learning and applications.
18. Robotic instrument segmentation sub-challenge part of the endoscopic vision challenge. https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/.
19. Islam M, Atputharuban DA, Ramesh R, Ren H. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. IEEE Robotics and Automation Letters. 2019; pp. 1–1. https://doi.org/10.1109/lra.2019.2900854.
20. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI. Automatic instrument segmentation in robot-assisted surgery using deep learning. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018.
21. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers; 2019.
22. Sarikaya D, Corso JJ, Guru KA. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection; 2017.
23. Chittajallu DR, Dong B, Tunison P, Collins R, Wells K, Fleshman J, Enquobahrie A. XAI-CBIR: explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019 https://doi.org/10.1109/isbi.2019.8759428.
24. Travis E, Woodhouse S, Tan R, Patel S, Donovan J, Brogan K. Operating theatre time, where does it all go? a prospective observational study. BMJ. 2014;15:349.
25. Twinanda AP, Yengera G, Mutter D, Marescaux J, Padoy N. RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE Transactions on Medical Imaging. 2018; pp.1–1. https://doi.org/10.1109/tmi.2018.2878055.
26. Ucuzal H, Arslan AK, Colak C. Deep learning based-classification of dementia inmagneticresonance imaging scans. 2019 International Artificial Intelligence and Data Processing Symposium(IDAP). 2019. https://doi.org/10.1109/idap.2019.887596.
27. Zhao Y, Zhao J, Zhao C, Xiong W, Li Q, Yang J. Robust Real-Time Object Detection Based on Deep Learning for Very High Resolution Remote Sensing Images. IGARSS 2019–2019. IEEE International Geoscience and Remote Sensing Symposium; 2019. https://doi.org/10.1109/igarss.2019.8897976.
28. Qu X, Wei T, Peng C, Du P. A fast face recognition system based on deepLearning. 2018 11th International Symposium on Computational Intelligence and Design (ISCID). 2018; https://doi.org/10.1109/iscid.2018.00072.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
30. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. 2017;29(9):2352–449.
31. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imag. 2017;36(1):86–97. https://doi.org/10.1109/tmi.2016.2593957.
32. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processingsystems; 2012. pp. 1097–1105.
33. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. https://doi.org/10.1109/cvpr.2016.90.
34. Deng J, Dong W, Socher R, Li L-J, Li Kai, Fei-Fei Li. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009. https://doi.org/10.1109/cvpr.2009.5206848.
35. Sahu M, Mukhopadhyay A, Szengel A, Zachow S. Tool and phase recognition using contextual cnn features. 2016.
36. Jo K, Choi Y, Choi J, Chung JW. Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. Appl Sci. 2019;9:2865.
37. Kanakatte A, Ramaswamy A, Gubbi J, Ghose A, Purushothaman B. "Surgical tool segmentation and localization using spatio-temporal deep network," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020; pp. 1658–1661. https://doi.org/10.1109/EMBC44109.2020.9176676.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.