Herag Arabian*, Firas Abou Dalla, Nour Aldeen Jalal, Tamer Abdulbaki Alshirbaji, and Knut Moeller

Attention Networks for Improving Surgical Tool Classification in Laparoscopic Videos

https://doi.org/10.1515/cdbme-2022-1172

Abstract: Deep learning approaches have been extensively developed to promote intelligent applications, such as surgical tool detection in surgical videos, inside the operating rooms (ORs). However, robustness and high-performance accuracy are demanding for such high-risk applications. In this paper, the Squeeze and Excitation (SE) and Convolutional Block Attention Module (CBAM) were employed and evaluated for improving surgical tool classification in laparoscopic videos. Experimental results explicate the advantage of both attention modules to the tool classification task. The SE and CBAM achieved mean average precision (mAP) of 88.38% and 88.40%, respectively, compared to 86.35% achieved by the base CNN model.

Keywords: Attention Modules, Deep Learning, Imbalanced Data, Laparoscopic Videos, Surgical Tool Classification.

1 Introduction

Artificial Intelligence (AI) has become a popular topic over the recent years. The implementation of deep learning, a subset of AI, has been adapted into everyday life through applications (APPs) of leisure and security. The ability and success of deep learning in image processing has paved the way for its utilization in different domains such as the medical field [1]. The use of AI during surgery has gained a lot of interest, opening a path towards smart operating rooms (ORs) that optimize surgical treatment and assist medical teams [1, 2].

Automatic recognition of surgical tools and phases [3, 4] in surgical videos, i.e. laparoscopic videos, is an important application of AI inside ORs of the future [5]. During last years, convolutional neural networks (CNNs) have been employed to perform tool classification in laparoscopic images [6–11]. To improve classification performance, various deep learning approaches have been developed and adapted to overcome the task-related challenges such as temporal relationship between adjacent frames and imbalanced data. In this respect, domain-related losses were introduced [7, 10], and temporal information along the video sequence was modelled using recurrent neural networks such as long short-term memory (LSTM) [6, 10, 12, 13].

Recently, attention modules were adopted into existing CNN architectures for finer and more informative feature focus. Shi et al. evaluated an attention driven model for real time tool detection, and they achieved state-of-the-art results on three datasets [11]. Similarly, Hu et al. evaluated the use of an attention guided technique with two Convolutional Neural Networks (CNN) achieving mean average precision of 86.90% on m2cai16-tool dataset. These works show that implementation of attention is beneficial and achieves better performance over base models.

In this study, the impact of attention modules on tool classification in laparoscopic videos was analysed. The CNN architecture of ResNet50 [14] is used, with pre-trained weights, as the base model. Two attention modules, the Squeeze and Excitation (SE) [15] and Convolutional Block Attention Module (CBAM) [16], added to the base model, were studied and compared. All models were evaluated on the Chlolec80 dataset using the average precision (AP). For further analysis, the prediction visualization technique of Gradient-weighted Class Activation Mapping (Grad-CAM) [17] was also performed.

***Corresponding author: Herag Arabian:** Institute of Technical Medicine (ITeM), Furtwangen University, Jakob Kienzle Str. 17, Villingen-Schwenningen, Germany, e-mail: H.Arabian@hs-furtwangen.de
Firas Abou Dalla: Furtwangen University, Villingen-Schwenningen, Germany
Nour Aldeen Jalal, Tamer Abdulbaki Alshirbaji: Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany; and Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany
Knut Moeller: Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany

2 Methods

To analyse the impact of attention modules on surgical tool classification in laparoscopic videos, three models were evaluated. The first model, hereby referred to as the base model, was based on the ResNet50 [14] architecture, due to its high performance on the similar task in previous work [6]. The Softmax layer was replaced with a sigmoid activation layer for multi-class classification. The second model incorporated the attention module of SE [15] into the base model, while the third involved the CBAM [16] attention module added to the base model. The three models were trained using the pre-trained parameters on ImageNet [18] as an initial start for the parameters, also known as transfer learning.

2.1 Model architecture & training options

Attention modules have been showing improvement in the representation abilities of CNN by finding correlations between the different feature spaces [15]. The first attention module implemented, the SE module, is able to improve neural network performance by calculating the different channel wise inter-dependencies. The second module selected was the CBAM, this attention block adds to the work of the SE by focusing not only on the channel information but also on the spatial domain [16]. These attention modules were chosen as they can be easily integrated into any existing architecture without significantly affecting computational performance.

Four attention modules were added into the existing base model at the early layers, inside the second and third convolution blocks of the ResNet50. The models were trained on 10 Epochs, at a varying learning rate starting at 0.002 and dropping 0.0009 after every iteration. An adaptive moment estimation (ADAM) optimization function with 0.9 gradient decay factor was used with a batch size of 50.

The implementation was carried out using MATLAB 2021a on a windows Intel Xeon 2.20 GHz with NVIDIA graphic card (GeForce RTX 2080Ti). The training time for the ResNet50 and ResNet50-SE was approximately 75 minutes per epoch. The ResNet50-CBAM required an additional 15 minutes per epoch. The Inference time on the testing set took approximately 17 minutes on each model.

2.2 Data description

The Cholec80 [9] database is a collection of 80 cholecystectomy procedures recorded at the University Hospital of Strasbourg. The recording was performed at a rate of 25 frames per second (fps) with tool annotations at 1 fps. The dataset contains seven surgical tools of Grasper, Hook, Bipolar, Scissors, Clipper, Irrigator, and Specimen Bag. The tools were annotated with the condition that at least half of the tool tip is present in the video frame.

2.3 Performance criteria

A performance criteria was set to analyse and compare all models. The first 40 videos from the Cholec80 dataset were used for training and the last 40 videos for testing. The performance of the models was evaluated based on the mean average precision (mAP).

To visualize the prediction focus area of the different models, the visualization technique of Grad-CAM [17] was utilised. The visualization was calculated between the Sigmoid activation layer and the Rectified Linear Unit (ReLU) layer "*activation_49_relu*" of the different models.

Table 1: Distribution of Tools in both Training and Testing Sets.

Tool	Training Set	Testing Set	Total
Grasper	56800	45788	102588
Bipolar	4106	4770	8876
Hook	48437	54669	103106
Scissor	1624	1630	3254
Clipper	3217	2769	5986
Irrigator	5384	4430	9814
Specimen Bag	5760	5702	11462

3 Results & Discussion

Table 1 presents the distribution of each tool in the training and testing sets. As seen from the distribution, the tools in each set are not uniformly distributed with a bias of the Grasper and Hook classes.

3.1 Model performance

The performance of the different models on the testing set are represented in Table 2. The results show that the attention modules improved the performance achieving, on average, an increase of around 2% over the base CNN model. This suggests that attention incorporated networks are able to focus on more informative features and neglect non-essential ones for tool classification.

Table 2: Mean Average precision (mAP) for the ResNet50, ResNet50-SE and ResNet50- CBAM models.

	ResNet50	ResNet50-SE	ResNet50-CBAM
mAP	86.35%	88.38%	**88.40%**

In Figure 1, the Average Precision (AP) of each tool and the mAP are shown. As seen from the results, the base architecture of ResNet50 achieved the lowest performance on the different tools when compared to the attention modules, except for Grasper. This exception was attributed to the high representation of the Grasper in the training dataset. This phenomenon can also be observed for the Hook class, where a near similar performance measure can be seen between the 3 models. This comes in contrast to the improvements achieved by the attention models for the other classes.

Both attention modules, SE and CBAM, achieved similar performance. However, the CBAM attention module showed a slight increase in performance compared to the SE. To get a better perspective and visual of where the model was focusing for decision making, the Grad-CAMs were computed and visualized.

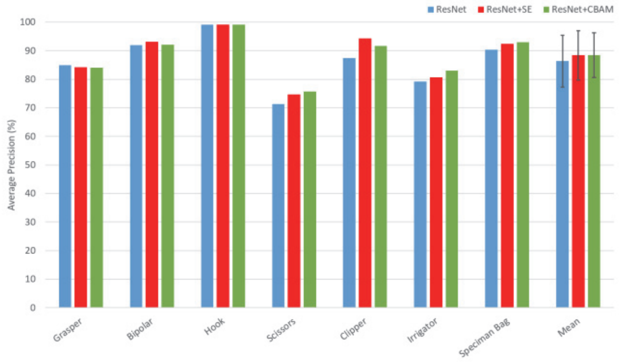

Figure 1: Average Precision (AP) of the tool presence detection of each tool for the three models.

Figure 2 represents the Grad-CAM visualizations on a sample of images for 2 classes of Clipper, Specimen Bag and Irrigator for each of the 3 models. As seen, the attention networks were able to focus more on the tool than the Base model thereby improving performance. A comparison between the Grad-CAMs of the SE module to the CBAM module showed that the area of concentration of the CBAM was larger than that of the SE. This influence is due to the module architecture concept, where the CBAM considers more information (channel and spatial information) than the SE.

The attention modules also helped addressing the imbalanced data problem faced. This is an important aspect as they were able to classify the tools with low incidences with high accuracy.

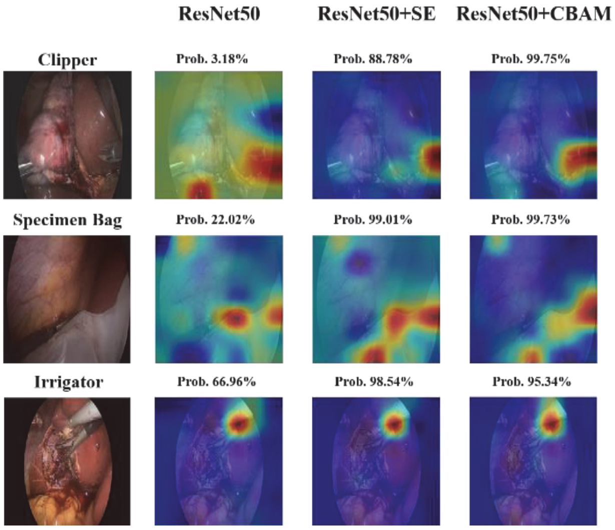

Figure 2: Grad-CAM visualizations and probabilities of 2 classes of Bipolar and Specimen Bag. Original image, ResNet50, ResNet50-SE and ResNet50-CBAM (Left to right).

4 Conclusion

In this study, the impact of adding attention modules on base CNN model performance for surgical tool classification was analysed. The results showed that the attention modules improved the performance of the models with a range of 2%. The Grad-CAM visualizations also showed that the attention modules helped the network to focus on more informative features. Attention modules were also able to overcome the imbalanced data problem. The CBAM architecture performed slightly better than the SE. Taking into consideration the focus area of the CBAM attention block, future work will rely on this attention module.

Author Statement

Research funding: This work was supported by the German Federal Ministry of Research and Education (BMBF) under project CoHMed/PersonaMed grant No. 13FH5I06IA and project CoHMed/DigiMedOP grant No. 13FH5I05IA. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent is not applicable. Ethical approval: The conducted research is not related to either human or animals use.

References

[1] Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P.: Surgical data science–from concepts toward clinical translation. Med. Image Anal. 76, 102306 (2022).

[2] Stauder, R., Ostler, D., Vogel, T., Wilhelm, D., Koller, S., Kranzfelder, M., Navab, N.: Surgical data processing for smart intraoperative assistance systems. Innov. Surg. Sci. 2, 145–152 (2017).

[3] Jalal, N.A., Alshirbaji, T.A., Möller, K.: Predicting surgical phases using CNN-NARX neural network. Curr. Dir. Biomed. Eng. 5, 405–407 (2019).

[4] Jalal, N.A., Alshirbaji, T.A., Docherty, P.D., Neumuth, T., Moeller, K.: A Deep Learning Framework for Recognising Surgical Phases in Laparoscopic Videos. IFAC-Pap. 54, 334–339 (2021).

[5] Lalys, F., Jannin, P.: Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. 9, 495–511 (2014). https://doi.org/10.1007/s11548-013-0940-5.

[6] Abdulbaki Alshirbaji, T., Jalal, N.A., Docherty, P.D., Neumuth, T., Möller, K.: A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos. Biomed. Signal Process. Control. 68, 102801 (2021). https://doi.org/10.1016/j.bspc.2021.102801.

[7] Alshirbaji, T.A., Jalal, N.A., Möller, K.: Surgical Tool Classification in Laparoscopic Videos Using Convolutional Neural Network. Curr. Dir. Biomed. Eng. 4, 407–410 (2018). https://doi.org/10.1515/cdbme-2018-0097.

[8] Hu, X., Yu, L., Chen, H., Qin, J., Heng, P.-A.: AGNet: attention-guided network for surgical tool presence detection. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. pp. 186–194. Springer (2017).

[9] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Trans. Med. Imaging. 36, 86–97 (2017). https://doi.org/10.1109/TMI.2016.2593957.

[10] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. 59, 101572 (2020). https://doi.org/10.1016/j.media.2019.101572.

[11] Shi, P., Zhao, Z., Hu, S., Chang, F.: Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network. IEEE Access. 8, 228853–228862 (2020).

[12] Alshirbaji, T.A., Jalal, N.A., Möller, K.: A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos. Curr. Dir. Biomed. Eng. 6, (2020).

[13] Jalal, N.A., Abdulbaki Alshirbaji, T., Docherty, P.D., Neumuth, T., Möller, K.: Surgical Tool Detection in Laparoscopic Videos by Modeling Temporal Dependencies Between Adjacent Frames. In: European Medical and Biological Engineering Conference. pp. 1045–1052. Springer (2020).

[14] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016).

[15] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018).

[16] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018).

[17] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017).

[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252 (2015).