Interpretable Deep Learning for Surgical Tool Management

Mark Rodrigues[1], Michael Mayo[1], and Panos Patros[2]

[1] Department of Computer Science, University of Waikato, Hamilton, New Zealand
[2] Department of Software Engineering, University of Waikato, Hamilton, New Zealand

Abstract. This paper presents a novel convolutional neural network framework for multi-level classification of surgical tools. Our classifications are obtained from multiple levels of the model, and high accuracy is obtained by adjusting the depth of layers selected for predictions. Our framework enhances the interpretability of the overall predictions by providing a comprehensive set of classifications for each tool. This allows users to make rational decisions about whether to trust the model based on multiple pieces of information, and the predictions can be evaluated against each other for consistency and error-checking. The multi-level prediction framework achieves promising results on a novel surgery tool dataset and surgery knowledge base, which are important contributions of our work. This framework provides a viable solution for intelligent management of surgical tools in a hospital, potentially leading to significant cost savings and increased efficiencies.

Keywords: Surgical tool dataset · multi-level predictions · hierarchical classification · surgery knowledge base.

1 Introduction

Surgical tool and tray management is recognized as a difficult issue in hospitals worldwide. Stockert and Langerman [16] observed 49 surgical procedures involving over two-hundred surgery instrument trays, and discovered missing, incorrect or broken instruments in 40 trays, or in 20% of the sets. Guedon et al. [5] found equipment issues in 16% of surgical procedures; 40% was due to unavailability of a specific surgical tool when needed. Zhu et al. [24] estimated that 44% of packaging errors in surgical trays at a Chinese hospital were caused by packing the wrong instrument, even by experienced operators. This is significant given the volumes; for example, just one US medical institution processed over one-hundred-thousand surgical trays and 2.5 million instruments annually [16].

There are tens of thousands of different surgical tools, with new tools constantly being introduced. Each tool differs in shape, size and complexity – often in very minor, subtle, and difficult to discern ways, as shown in Fig.1. Surgical sets, which can contain 200 surgical tools, are currently assembled manually

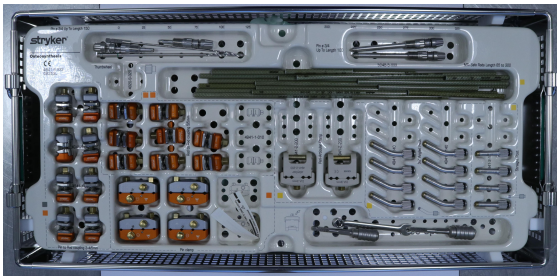

Fig. 1. Surgical tools - Hoffman Compact instruments and implants

[11] but this is a difficult task even for experienced packing technicians. Given that surgical tool availability is a mission-critical task, vital to the smooth functioning of a surgery, ensuring that the tool is identified accurately is extremely important. Al Hajj et al. [2] reviewed convolutional neural network (CNN) architectures and a range of imaging modalities, applications, tasks, algorithms and detection pipelines used for surgical segmentation. They pointed out that hand crafted and hand engineered features had also been used for this task, and Bouget et al. [3] reviewed predominant features used for object-specific learning with surgical tools, and listed colour, texture, gradient and shape as being important for detection and classification. Yang et al. [19] presented a review of the literature regarding image-based laparoscopic tool detection and tracking using CNNs, including a discussion of available datasets and CNN-based detection and tracking methods. While CNNs can therefore provide viable solutions for surgical tool management, understanding how the CNN makes a prediction is important for building trust and confidence in the system.

Interpretibility of predictions is then a critical issue – Rudin et al. [12] stated that interpretable machine learning is about models that are understood by humans, and interpretability can be achieved via separation of information as it traverses through the CNN models. Zhang et al. [21] developed an interpretable model that provided explicit knowledge representations in the convolutional layers (conv-layers) to explain the patterns that the model used for predictions. Linking middle-layer CNN features with semantic concepts for predictions provided interpretation for the CNN output [15, 22, 23]. How mid-level features of a CNN represent specific features of surgical tools and how they can provide hierarchical predictions is the focus of our work. CNNs learn different features of images at different layers, with higher layers extracting more discriminative features [20]. By associating feature maps at different CNN levels to levels in a hierarchical tree, a CNN model could incorporate knowledge of hierarchical categories for better classification accuracy. The model developed by Ferreira et al. [4] addressed predictions across five categorisation levels: gender, family, category, sub-category and attribute. The levels constituted a hierarchical structure, which was incorporated in the model for better predictions. The benefit of this hierarchical and interpretable approach for surgical tool management is

that end users can then make rational, well reasoned decision on whether they can trust the information presented to them [12].

Wang et al. [18] discussed an approach to fine tuning that used wider or deeper layers of a network, and demonstrated that this significantly outperformed the traditional approaches which used pre-trained weights for fine-tuning. Going deeper was accomplished by constructing new top or adaptation layers, thereby permitting novel compositions without needing modifications to the pre-trained layers for a new task. Shermin et al. [14] showed that increasing network depth beyond pre-trained layers improved results for fine-grained and coarse classification tasks. We build on these approaches in our multi-level predictor.

Table 1. Surgical Datasets

Characteristic	CATARACTS	Cholec80	Surgical Tools
Size or Instances	50 videos	80 Videos	18300 images
Database Focus	Cataract Surgeries	Cholecystectomy Surgeries	Orthopaedics and General Surgery
Type of Surgery	Open Surgery	Laparoscopic	Open Surgery
Default Task	Detection	Detection	Classification
Type of Item	Videos	Videos	RGB Images
Number of Classes	21	7	361
Images Background	Tissue	Tissue	Flat colours
Image Acquisition Platform / Device	Toshiba 180I camera and MediCap USB200 recorder	Not Specified	Canon D-80 Camera and Logitech 922 Pro Stream Webcam
Image Illumination	Microscope Illumination	Fibre-optic in-cavity	Natural Light, LED, Fluorescent
Distance to Object	V.Close - Microscope	Close - in-cavity	30-cms to 60-cms
Annotations	Binary	Bounding Boxes	Multiple level
Dataset Organisation	500,000 frames each in Training and Test Sets	86,304 & 98,194 frames in Training and Test Set	14,640 images in Training and 3,660 in Validation set
Structure	Flat	Flat	Hierarchical
Image Resolution	1920x1080 pixels	Not Specified	600 x 400 pixels

2 Surgical Tool Dataset Overview

Kohli et al. [7] and Maier-Hein et al. [10] discussed the problems faced by the machine learning community stemming from a lack of data for medical image evaluation, which significantly impairs research in this area. There is just not enough high quality, well annotated data, representative of the particular surgery – a shortfall that needs to be addressed. Most medical datasets are one-off solutions for specific research projects, with limited coverage and restricted in numbers of images or data points [10]. To address this, we plan to create and curate a surgical tool dataset with tens of thousands of tool images across all surgical specialities with high quality annotations and reliable ground-truth information.

Since surgery is organised along specialities, each with its own categories, a hierarchical classification of surgical tools would be extremely valuable. We therefore developed our initial surgical dataset with a hierarchical structure based on the surgical speciality, pack, set and tool. We captured RGB images of surgical tools using a DSLR camera and a webcam and tried to provide consideration to achieving viewpoint invariant object detection with different backgrounds, illumination, pose, occlusion and intra-class variations captured in the images. We focused on two specialities – Orthopaedics and General Surgery – of the 14 specialities reported by the American College of Surgeons [1]. The former offers a wide range of instruments and implants, while the latter covers the most common surgical tools. We propose to add the other specialities in a phased manner, and will make the dataset publicly available to facilitate research in this area.

CNNs have been successfully used for the detection, segmentation and recognition of surgical tools [9]. However, the datasets currently available for surgical tool detection present very small instrument sets; to illustrate this, the Cholec80, EndoVis 2017 and m2cai16-tool datasets have seven instruments, the CATARACTS dataset has 21 instruments, the NeuroID dataset has eight instruments and the LapGyn4 Tool Dataset has three instruments [2, 17]. While designing CNNs to recognise seven or eight instruments for research purposes may be justifiable, this is nowhere nearly adequate enough for real work conditions. Any model trained using this data is unlikely to be usable anywhere else, not even in the same hospital six months later. We needed to develop a new dataset for our work as these surgical tool datasets did not offer a sufficiently large variety or number of tools for analysis, nor were they arranged hierarchically. A comparison of our dataset with CATARACTS [2] and Cholec80 [17], two important publicly available datasets, is presented in Table 1.

Table 2. Surgery Knowledge Base (Excerpt)

Speciality	Pack	Set	Tool
Orthopaedics	VA Clavicle Plating Set	LCP Clavicle Plates	Clavicle Plate 3.5 8 Hole Right
Orthopaedics	Trimed Wrist Fixation System	Fixation Fragment Specific	Dorsal Buttress Pin 26mm
General Surgery	Cutting & Dissecting	Scissors	9 Metzenbaum Scissors
General Surgery	Clamping & Occluding	Forceps	6 Babcock Tissue Forceps

2.1 Surgery Knowledge Base

Setti [13] points out that most public benchmark datasets only provide images and label annotations, but providing additional prior knowledge can boost performance of CNNs. To complement the dataset, we developed a comprehensive surgery knowledge-base (Table 2) as an attribute-matrix which makes rich information available to the training regime. This proved to be a convenient and

useful data structure that captures rich information of class attributes – or the nameable properties of classes – and makes it readily available for computational reasoning [8]. We developed the knowledge representation structure for 18,300 images to provide rich, multi-level and comprehensive information about each image. The attribute matrix data structure proved to be easy to work with, simple to change and update, and it also provided computational efficiencies.

3 Experimental Method

We implemented our project in Tensorflow v-2.4.1 and Keras v-2.4.3. Our architecture consists of a ResNet50V2 network [6] which we trained on the Surgical Tool training dataset, by replacing the top layer with a dropout and dense layer with 361 outputs. We initially did not use the knowledge base annotations, only the tool labels and trained with the configuration in Table 4 with early stopping on validation categorical accuracy. We were able to obtain good predictions from this model with accuracy score at 93.51%, but only at the tool level. We then used this pre-trained architecture with surgical tool weights as our base model, froze the base model, and added separate classification pipelines, one for each prediction of interest - speciality, set, pack and tool (See Fig. 2). We relied on the knowledge base annotations which provided data for two specialities, twelve packs, thirty-five sets and 361 possible tools, and used it to create data-frames for the training and validation data. Each image was associated with the relevant annotations for each output, in the form of columns of text values or categorical variables representing the multiple classes for each output. This multi-task framework effectively shared knowledge of the different attribute categories for each image or visual representation. We developed a custom data handler for the training data (x_set) and for the labels for each of the four outputs (y_cat, y_pack, y_set, y_tool), and used one hot encoding to represent the categorical variables in our model. We then implemented training and validation data generators based on our custom data handler to provide batches of data to the model. Our model was compiled with one input (image) and four outputs.

Table 3. Results - Val accuracy with output at different layers

All Outputs at:	Total Parameters	Parameters Trained	Speciality	Pack	Set	Tool
Conv2_block1_1_relu	700,570	686,490	0.956	0.356	0.258	0.091
Conv3_block1_1_relu	1,210,266	948,634	0.989	0.621	0.507	0.231
Conv4_block1_1_relu	3,060,634	1,472,922	0.997	0.927	0.851	0.663
Conv5_block1_1_relu	11,625,370	2,521,498	0.999	0.975	0.945	0.890

We tested outputs at different layers to evaluate the impact of changing the depth of the network, with the results in Table 3. In each experiment, parameters available and actually trained were controlled by adjusting the numbers of layers. An operation within a block in ResNet50V2 consisted of applying convolution, batch normalisation and activation to an input; we obtained our outputs after

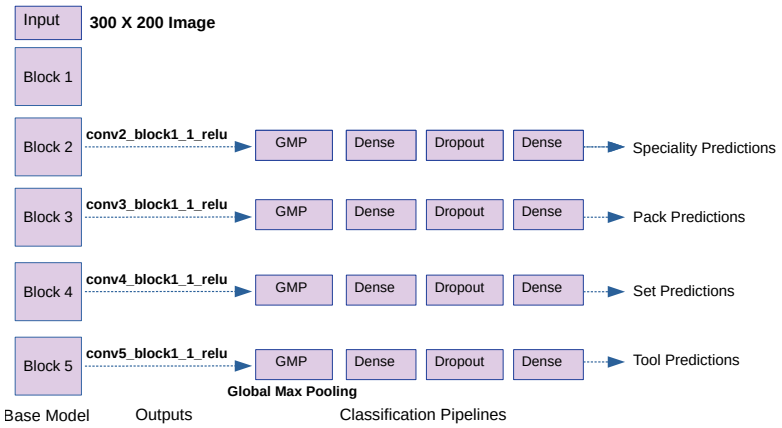

Fig. 2. Resnet50V2 Architecture with Multiple Outputs

the first operation in each block. These outputs were fed to external global max pooling and dense layers. A dropout layer regulated training – we replaced this with a batch normalisation layer but results did not improve. Since this was a multi-class problem, a dense layer with softmax activation was used for the final classification of each prediction, customised to the relevant number of classes. As we expected, better results were obtained by including more layers and by training more parameters – best results were obtained by including all layers up to Block 5. However, it is noteworthy that high accuracy was obtained for specific predictions even early in the model – for example, predictions for speciality were at 95.6% by block 2, for pack and set were at 92.7% and 85.10% at block 4 and for tool at 89% at block 5. Clearly it was possible to obtain accurate predictions for higher level categories using early layers of the model. This is explored further with the objective of improving interpretability for the end user, while reducing the total number of parameters that needed to be trained in the model.

Table 4. Training Configuration

Parameter	Optimiser	Learning Rate	Batch Size	Activation	Loss	Metric
Value	Adam	0.001	64	Softmax	Categorical Crossentropy	Categorical Accuracy

The training set images from the surgery dataset and annotations from the knowledge base were used for training, with real time training data augmentation – including horizontal flip, random contrast and random brightness operations. We used the configuration in Table 4, the initial learning rate of 0.001 was decreased to 0.0001 at epoch 45 and to 0.00005 at epoch 75. A dropout rate of

0.2 was imposed. We implemented early stopping on val loss with a patience of 20 epochs. The total parameters in the model were 10,511,258, and parameters trained were 1,407,386 in each of the experiments.

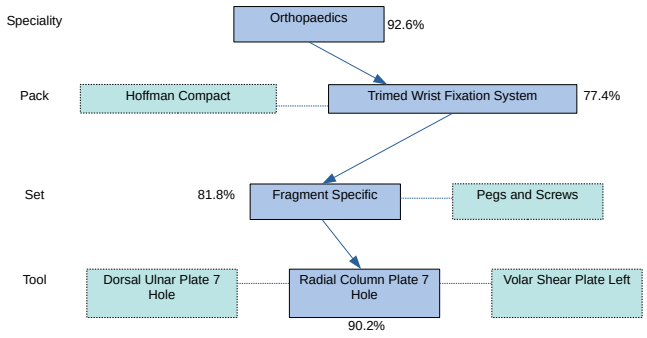

Fig. 3. Interpretable multi-level predictions

1. ImageNet Training: For an initial baseline experiment, we used a ResNet50V2 model with ImageNet weights and four separate classification outputs were trained, one for each hierarchy – speciality, set, pack and tool.
2. Surgical Tool Training: We used the pre-trained base model with surgical tool weights, and trained the model with its four classification pipelines using the configuration as in Table 4 and architecture as in Fig. 2.
3. Depth Adjusted Surgical Tool Training: We used the pre-trained model with surgical tool weights as before, but changed the levels within the blocks of the ResNet-50V2 model from which we obtained outputs, thereby adjusting the depth of training. The outputs from Block 5 and 2 were obtained from conv"x"_block1_1, and from Block 3 and 4 were from conv"x"_block4_2. We did this to evaluate the effects of changing depths on the prediction accuracy; this was a minor change within the block but the total number of parameters trained were maintained the same.

4 Results and Conclusions

Our results, on a separate test subset of data, are shown in Table 5. The test data was images that the model had not seen before, as a sample of 400 random images across all classes had been reserved for testing. Training with ImageNet weights did not provide good results, but the use of surgical tool weights demonstrated that the model had captured relevant information about the dataset and was able to provide good predictions at multiple levels. In this architecture, by extracting multiple predictions along layers from coarse to fine as data traverses the CNN, early layers provided predictions corresponding to specialities while

Table 5. Architecture Results - Macro score or average for all classes

Level	Metric	ImageNet	Surgical-Tools	Surgical-Tools Depth Adjusted
Speciality score	Accuracy score	0.90	0.94	0.94
	Hamming Loss	0.10	0.06	0.06
	f1 Score	0.73	0.84	0.83
	Precision score	0.93	0.95	0.95
	Recall score	0.96	0.99	0.99
Pack score	Accuracy score	0.41	0.63	0.77
	Hamming Loss	0.59	0.37	0.23
	f1 Score	0.25	0.53	0.73
	Precision score	0.43	0.67	0.76
	Recall score	0.30	0.55	0.73
Set score	Accuracy score	0.31	0.84	0.89
	Hamming Loss	0.69	0.16	0.11
	f1 Score	0.24	0.79	0.84
	Precision score	0.36	0.82	0.85
	Recall score	0.25	0.80	0.87
Tool score	Accuracy score	0.20	0.90	0.90
	Hamming Loss	0.80	0.10	0.10
	f1 Score	0.16	0.86	0.86
	Precision score	0.78	0.91	0.91
	Recall score	0.27	0.91	0.90

later layers provide finer predictions, such as tool classifications (Fig. 3). It was easy for the CNN to distinguish between our two speciality classes, since General Surgery tools are visually different from orthopaedic tools – as we add more specialities where the visual distinction is not so clear, we may need to train at deeper levels. As the classes increased to 12, 35 and 361 for pack, set and tool respectively, predictions from deeper layers were needed. These hierarchical predictions are expected to provide better interpretability since multiple predictions can be tested and evaluated against each other for consistency or error by the end user. Adjusting the depths of layers used as outputs for predictions improved the results, even within the same block, demonstrating that more features are learned as the data travels through the CNN layers.

We developed a CNN framework that successfully utilised the hierarchical nature of surgical tool classes to provide a comprehensive set of classifications for each tool. This framework was deployed and tested on a new surgical tool dataset and knowledge base. The multi-level prediction system provides a good solution for classification of other types of medical images, if they are hierarchically organised with a large number of classes.

References

1. ACS. What are the surgical specialties? https://www.facs.org/education/resources/ medical-students/faq/specialties, 2021. Accessed: 15/2/2021.
2. H. Al Hajj, M. Lamard, P. H. Conze, S. Roychowdhury, X. Hu, G. Marsalkaite, and M. Sahu. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis*, 52:24–41, 2019.
3. D. Bouget, M. Allan, D. Stoyanov, and P. Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical Image Analysis*, 35:633, 2017.
4. Beatriz Ferreira, Luis Baia, Joao Faria, and Ricardo Sousa. A unified model with structured output for fashion images classification. In *AI for Fashion - The third international workshop on Fashion and KDD, London, United Kingdom*, 2018.
5. A.C. Guedon, L.S. Wauben, A.C. van der Eijk, A. S. Vernooij, F.C. Meeuwsen, M. van der Elst, V. Hoeijmans, J. Dankelman, and J. J. van den Dobbelsteen. Where are my instruments? Hazards in delivery of surgical instruments. *Surgical endoscopy, 30(7).*, 2016.
6. K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC)*. IEEE Computer Society, 2016.
7. Marc D. Kohli, Ronald M. Summers, and J. Raymond Geis. Medical image data and datasets in the era of machine learning – white paper from the 2016 C-MIMI Meeting Dataset Session. *Journal of Digital Imaging (2017) 30.*, 2017.
8. C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 453-465*, 2014.
9. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature.*, 10(1038): 436–44, 2015.
10. L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. Marz, and et al. Surgical data science - from concepts to clinical translation. *ArXiv, abs/2011.02284*, 2020.
11. J. M. Mhlaba, E. W. Stockert, M. Coronel, and A. J. Langerman. Surgical instrumentation: The true cost of instrument trays and a potential strategy for optimization. *Journal of Hospital Administration*, 4:6, 2015.
12. Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 Grand Challenges. *ArXiv, abs/2103.11251.*, 2021.
13. F. Setti. To know and to learn – about the integration of knowledge representation and deep learning for fine-grained visual categorization. In *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2018.
14. Tasfia Shermin, Manzur Murshed, Shyh Teng, and Guojun Lu. Depth augmented networks with optimal fine-tuning. *ArXiv, abs/1903.10150*, 2019.

15. Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *International Conference on Computer Vision (ICCV)*, 2015.
16. E. W. Stockert and A. J. Langerman. Assessing the magnitude and costs of intraoperative inefficiencies attributable to surgical instrument trays. *Journal of the American College of Surgeons*, 219(4):646–655, Oct 2014.
17. A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36, 2017.
18. Yu-Xiong Wang, D. Ramanan, and M. Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
19. Congmin Yang, Zijian Zhao, and Sanyuan Hu. Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature. *Computer Assisted Surgery, 25:1, 15-28*, 2020.
20. M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham*, 2014.
21. Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, and S. C. Zhu. Interpretable cnns for object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
22. Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting CNNs via decision trees. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
23. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene CNNs. In *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA,*, 2015.
24. X. Zhu, L. Yuan, T. Li, and P. Cheng. Errors in packaging surgical instruments based on a surgical instrument tracking system: an observational study. *BMC Health Services Research, 19:176*, 2019, 2019.