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 for multi-level classification of surgical tools, with a set of property predictions. Predictions are obtained from multiple levels of the model, and high accuracy is obtained by adjusting the depth of layers selected for predictions. Our architecture improves interpretability by providing a comprehensive set of predictions for each tool, allowing users to make rational decisions about whether to trust the model based on multiple pieces of information. These predictions can be evaluated against each other for consistency and error-checking. Important contributions of our work are the interpretable multi-level architecture, a novel surgical tool dataset, and a surgery knowledge base. This architecture 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 [14] 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. [4] found equipment issues in 16% of surgical procedures; 40% was due to unavailability of a specific surgical tool when needed. Zhu et al. [21] 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 [14].

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 [10] but this is a difficult task even for experienced packing technicians. Given

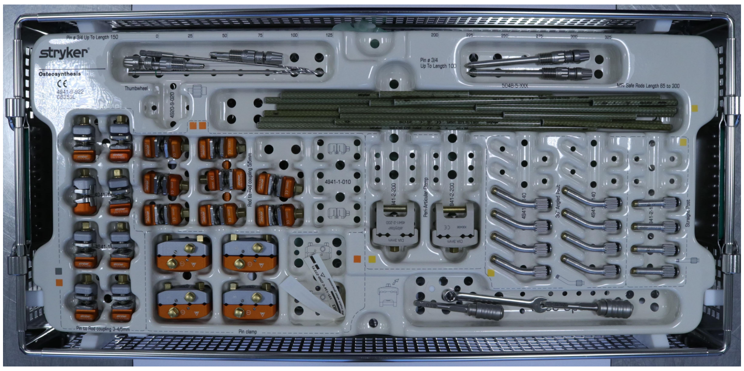

Fig. 1. Surgical tools - Hoffman Compact instruments and implants

that surgical tool availability is a mission-critical task, vital to the smooth functioning of a surgery, ensuring that the tool is identified accurately and in an understandable manner is extremely important.

Interpretibility of predictions therefore is a critical issue – Rudin et al. [11] 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. [18] 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 [20] [13] [19]. The questions around how mid-level features of a CNN represent specific features of surgical tools and how they can be used to 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 [17]. By associating feature maps at different CNN levels to the levels in a hierarchical tree, a CNN model could incorporate knowledge of hierarchical categories and relationships for better classification accuracy. The model developed by Ferreira et al. [3] 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 heirarchical 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 [11].

Wang et al. [16] discussed an approach to fine tuning CNNs that used wider or deeper layers of a network, and demonstrated that this significantly outperformed traditional approaches that 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 pre-trained layers for a new task. Shermin et al. [12] 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.

2 Surgical Tool Dataset Overview

We developed our 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 to create the initial dataset. We focused on two specialities – Orthopaedics and General Surgery – out of the 14 specialities reported by the American College of Surgeons [1]. The first of these specialities offers a wide range of instruments, implants and screws, while the second speciality covers common instruments used across all open surgery.

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 in Training and 500,000 frames in Test Set	86,304 frames in Training and 98,194 in 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

CNNs have been successfully used for the detection, segmentation and recognition of objects in images, including surgical tools detection [8]. 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] [15]. While designing and testing CNNs to recognise seven or eight instruments for research purposes may be justifiable, this is entirely inadequate for real work conditions. Any model trained using such small datasets is unlikely to be usable anywhere else, not even in the same hospital six months later. We needed to develop a new surgical tool dataset that provided a large variety and number of tools for analysis, and which was arranged hierarchically. A comparison of our dataset with CATARACTS [2] and Cholec80 [15], two important publicly available datasets, is presented in Table 1.

Kohli et al. [6] and Maier-Hein et al. [9] 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 – and this is a shortfall that needs to be addressed. Currently, most of the medical datasets are one-off solutions for specific research projects, with limited coverage and are restricted in terms of size to hundreds – rather than thousands – of images or data points [9]. We therefore plan to create and curate a large surgical tool dataset of 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.

Table 2. Surgery Knowledge Representation (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	Trimed Wrist Fixation Fragment Specific	Dorsal Buttress Pin 26mm
General Surgery	Cutting and Dissecting	Scissors	9 Metzenbaum Scissors
General Surgery	Clamping and Occluding	Forceps	6 Babcock Tissue Forceps

2.1 Surgery Knowledge Base

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 [7]. 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.

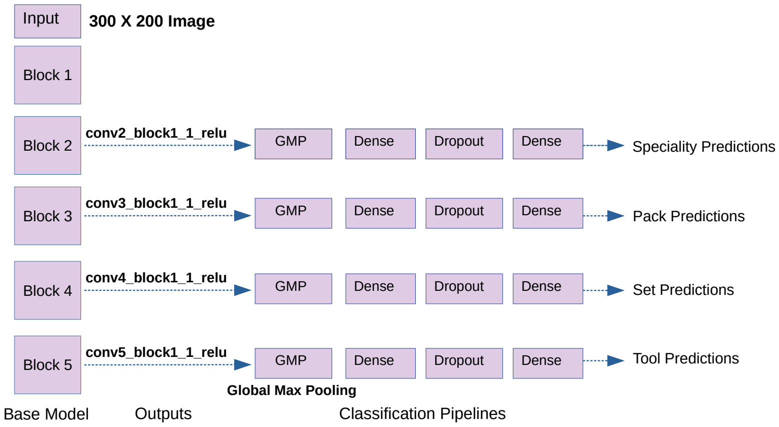

Fig. 2. Resnet50V2 Architecture with Multiple Outputs

3 Experimental Method

We implemented our system in Tensorflow v-2.4.1 and Keras v-2.4.3. We created data-frames for the training and validation data, with paths to the images and the annotations for each of the multiple outputs in the form of columns for each output. The dataset annotations contained categorical variables in the form of text values, representing the multiple classes for each output. We used one hot encoding to represent the categorical variables in a suitable format for our model. Our model was designed to provide four multi-level outputs for each image input, and we developed a custom data handler to provide the training data (x set) along with labels for each of the four outputs (y cat, y pack, y set, y tool). We then used train and validation data generators based on our custom data handler to provide batches of data to the model.

Our architecture consists of a ResNet50V2 network [5] as base network. We initialised the model using weights obtained by pre-training the ResNet50V2 model on the Surgical Tool test dataset. The base network was then frozen for all experiments. We added separate classification pipelines to the base network, one for each prediction of interest - speciality, set, pack and tool (See Fig. 2). Prediction pipelines were built by obtaining outputs from the activation layers at specific blocks. We did not get good results with global average pooling, but a global max pooling layer and dense layer formed an effective pipeline. We used categorical cross-entropy as the loss and categorical accuracy as the metric for each output provided the best results. The model was compiled with one input (image) and four outputs.

We tested outputs at different layers to evaluate the impact of changing the depth of the network, and our results are presented in Table 3. In each experiment, all the predictions were obtained from the same level of the Resnet50V2

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

All Outputs at:	Total Parameters	Parameters Trained	Speciality	Pack	Set	Tool
Conv2	700,570	686,490	0.956	0.356	0.258	0.091
Conv3	1,210,266	948,634	0.989	0.621	0.507	0.231
Conv4	3,060,634	1,472,922	0.997	0.927	0.851	0.663
Conv5	11,625,370	2,521,498	0.999	0.975	0.945	0.890

model, and both the total number of parameters available to train and the numbers of layers included in the model. Outputs were obtained early in each block (conv"X"_block1_1_relu). An operation within a block in ResNet50V2 consists of applying convolution, batch normalisation and activation to an input, and we obtain our outputs after the first operation in each block. These outputs were fed to the external global max pooling and dense layers. A dropout layer was used to regulate training – replacement with a batch normalisation layer did not improve results. A dense layer with softmax activation was used for the final classification of each multi-class 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 of the ResNet-50V2 model. 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 disentangle information as it traverses the CNN and to obtain predictions for higher level categories using early layers of the model. This is explored further in our next set of experiments 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
Value	Adam	0.001	64	Softmax	Categorical Crossentropy

Our prototype system was trained on the surgery dataset and knowledge base, which captured two specialities, twelve packs, thirty-five sets and 361 possible tools. Real time training data augmentation was conducted on the test set, including horizontal flip, random contrast and random brightness operations. We experimented with SGD but finalised on the configuration as shown in Table 4. The initial learning rate of 0.001 was decreased to 0.00001 at epoch 45 and

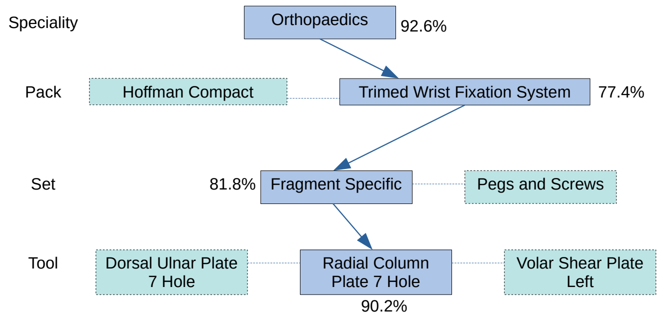

Fig. 3. Interpretable multi-level predictions

to 0.000005 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.

1. ImageNet Training: We trained the model with the four classification pipelines with ImageNet weights with early stopping implemented for validation loss. The base model was frozen, and the four separate classification outputs were trained, one for each hierarchy – speciality, set, pack and tool.
2. Surgical Tool Training: We pre-trained the ResNet50V2 model on the Surgical Tool test dataset, by replacing the top layer with a dropout and dense layer with 361 outputs. We used only the tool labels with the training configurations as in Table 4 with early stopping on validation loss. We saved this model after training and used it as the base model. We froze the base model, and trained the model with its four classification pipelines.
3. Depth Adjusted Surgical Tool Training: We used the pre-trained surgical tool weights, but changed the levels within the blocks of the ResNet50V2 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_relu, and from Block 3 and 4 were from conv"X"_block4_2_relu. We did this to evaluate the effects of changing depths within the CNN on the prediction accuracy; this was a minor change within the block but the total number of parameters trained were controlled and 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

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

Level	Metric	ImageNet	Surgical-Tools	Surgical-Tools Depth Adjusted
Speciality	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	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	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	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

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 later layers provide finer predictions, such as tool classifications. 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. It was visually easy for the CNN to distinguish between our two speciality classes, since General Surgery tools are significantly different visually 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 number of classes increased to 12, and 35 and 361 for pack, set and tool respectively, predictions from deeper layers were needed. These hierarchical predictions can provide better interpretability since multiple predictions – as presented in Fig. 3 – can be tested and evaluated against each other for consistency or error by the end user. Since the user has multiple pieces of information, decomposed into sub-parts or sub-predictions, they are in a better position to make trust based, well informed final decisions.

The multi-level model provides a simple and practical solution for surgical tool management by capturing and presenting relevant predictions as information travels through the CNN. This multi-level prediction system can provide 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. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Marsalkaite, G., Sahu, M.: Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical Image Analysis 52, 24–41 (2019)
3. Ferreira, B., Baia, L., Faria, J., Sousa, R.: A unified model with structured output for fashion images classification. ArXiv abs/1806.09445 (2018)
4. Guedon, A., Wauben, L., van der Eijk, A., Vernooij, A.S., Meeuwsen, F., van der Elst, M., Hoeijmans, V., Dankelman, J., van den Dobbelsteen, J.J.: Where are my instruments? hazards in delivery of surgical instruments. Surgical endoscopy, 30(7). (2016)
5. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC). IEEE Computer Society (2016)
6. Kohli, M.D., Summers, R.M., Geis, J.R.: Medical image data and datasets in the era of machine learning – whitepaper from the 2016 c-mimi meeting dataset session. Journal of Digit Imaging (2017) 30. (2017)
7. Lampert, C.H., Nickisch, H., Harmeling, S.: 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)
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 10(1038), 436–44 (2015)
9. Maier-Hein, L., Eisenmann, M., Sarikaya, D., Marz, K., et al.: Surgical data science - from concepts to clinical translation. ArXiv, abs/2011.02284 (2020)
10. Mhlaba, J.M., Stockert, E.W., Coronel, M., Langerman, A.J.: Surgical instrumentation: The true cost of instrument trays and a potential strategy for optimization. Journal of Hospital Administration 4, 6 (2015)
11. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. ArXiv, abs/2103.11251. (2021)
12. Shermin, T., Murshed, M., Teng, S., Lu, G.: Depth augmented networks with optimal fine-tuning. ArXiv, abs/1903.10150 (2019)
13. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: ICCV (2015)
14. Stockert, E.W., Langerman, A.J.: 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)

15. 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, 1 (2017)
16. Wang, Y.X., Ramanan, D., Hebert, M.: Growing a brain: Fine-tuning by increasing model capacity. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham (2014)
18. Zhang, Q., Wang, X., Wu, Y.N., Zhou, H., Zhu, S.C.: Interpretable cnns for object classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
19. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting cnns via decision trees. arXiv:1802.00121 (2019)
20. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. arXiv:1412.6856 (2015)
21. Zhu, X., Yuan, L., Li, T., Cheng, P.: Errors in packaging surgical instruments based on a surgical instrument tracking system: an observational study. BMC Health Serv Res 2019, 19 (2019)