Psychomotor Skills Assessment via Human Experts, Simulators, and Artificial Intelligence

Roger Smith, Danielle Julian

AdventHealth Nicholson Center

Celebration, FL

Roger.Smith@AdventHealth.com, Danielle.Julian@AdventHealth.com

ABSTRACT

Surgical education programs present significant challenges for the creation of accurate and executable assessment methods and metrics for psychomotor skills. This assessment has historically been accomplished by (1) observation and scoring by a qualified expert, (2) examination and scoring of a finished product, or (3) simulator data collection and mathematic scoring of performance. Each of these offer unique advantages and limitations during real execution.

Surgical courses at the Nicholson Center combine all three of these assessment methods, while also investigating new technologies that offer improvements. Given the advancements that are being reported for artificial intelligence and deep learning we have studied the applicability of these techniques to the assessment of surgical skills. Videos of student performance during simulated exercises were segmented for analysis by a deep neural network.

This research project explored the applicability of existing cloud-based neural network systems to analyze the contents of simulated surgical videos, identify the objects in the scene, identify the activity that was being performed, and apply a score to the quality of performance demonstrated. This paper describes the first objective of this experiment. Companies with expertise in creating neural networks are making these systems available to customers as cloud-based services, with supporting consulting services and training materials. This project made use of these services rather than attempting to build a neural network and hosting service from scratch. As offered, without modification, these services are capable of accurately identifying thousands of objects that are commonly found in online videos. The networks can also be retrained to recognize a custom library of objects. We successfully retrained two of these networks to identify the objects that appear in simulated surgical videos. This customized network could potentially become a web-based service for use by other surgical training centers.

ABOUT THE AUTHORS

Roger Smith, Ph.D., has spent 25 years creating leading-edge simulators for the Department of Defense and Intelligence agencies. He is currently the Chief Technology Officer for the AdventHealth Nicholson Center where he is responsible for establishing technology strategy and leading research experiments. He has served as the CTO for the U.S. Army PEO-STRI; VP and CTO for training systems at Titan Corp; and VP of Technology at BTG Inc. He holds a Ph.D. in Computer Science, a Doctorate in Management, an M.S. in Statistics, and a B.S. in Applied Mathematics. He has published 3 professional textbooks on simulation, 12 book chapters, and over 100 journal and conference papers. His most recent book is *A CTO Thinks About Innovation*. He has served on the editorial boards of the *Transactions on Modeling and Computer Simulation* and the *Research Technology Management* journals.

Danielle Julian, M.S., is a Senior Research Scientist at AdventHealth's Nicholson Center. Her current research focuses on robotic surgery simulation and effective surgeon training. Her current projects include intelligent tutoring system, rapid prototyping of surgical education devices, and the evaluation of robotic simulation systems. She is a certified instructor for surgical robotics courses delivered to surgeons and OR staff members. Her background includes research in Human Factors and learning and training to enhance the higher-order cognitive skills of military personnel. She is currently a Ph.D. student in Modeling and Simulation at the University of Central Florida where she previously earned an M.S. in Modeling and Simulation, Graduate Simulation Certificate in Instructional Design, and a B.S. in Psychology.

Psychomotor Skills Assessment via Human Experts, Simulators, and Artificial Intelligence

Roger Smith, Danielle Julian

AdventHealth Nicholson Center

Celebration, FL

Roger.Smith@AdventHealth.com, Danielle.Julian@AdventHealth.com

BACKGROUND

Surgical education programs present significant challenges for the creation of accurate and executable assessment methods and metrics for psychomotor skills. This assessment has historically been accomplished by (1) observation and scoring by a qualified expert, (2) human examination and scoring of a finished product, or (3) simulator data collection and mathematic scoring of performance. Each of these offer unique advantages and limitations when coupled with educational programs.

(a) dV-Trainer Simulator

(b) dV-Trainer Exercise Example

(c) dV-Trainer Metrics

Global Evaluative Assessment of Robotic Skills (GEARS)				
Depth Perception				
1	2	3	4	5
Bimanual Dexterity				
1	2	3	4	5
Efficiency				
1	2	3	4	5
Force Sensitivity				
1	2	3	4	5
Robotic Control				
1	2	3	4	5

(d) GEARS Subjective Metrics

Figure 1. Robotic Surgery Training Simulator and Assessment Metrics.,

Surgical education programs around the world have relied upon and been limited by these traditional assessment methods. Students in medical school, residency, and continuing medical education programs all require assessments of their performance to determine whether they have mastered the necessary skills (Tian, 2007, Davis 1999, 2006). For centuries, surgeons have been evaluated by their mentors using very subjective criteria and standards which varied widely from one institution to another. In the modern age of standardization, these processes have received some form of commonality through the creation and validation of metric tools that can be used in the same way across many organizations (Martin, 1997; van Hove, 2010; Gho, 2012). However, the use of standard metrics is limited by the lack of standardization in medical education and by the proliferation of multiple tools for the same activities.

More recently, computerized simulators, with the ability to accurately measure each activity, have opened the door for a new form of assessment (Figure 1). Companies that produce these simulators work with surgical researchers to arrive at benchmarks that indicate different levels of proficiency. Educational programs that use these simulators are then equipped with identical metrics that are objective assessments of the activities performed (Abboudi et al, 2013). This standardization is extremely useful in moving medical education into a more performance-based system.

Problems with Traditional Metrics

Though the existing methods can be used very effectively, each presents its own challenges that can significantly limit the effectiveness of surgical education programs.

Using human experts to evaluate significant numbers of repeated educational activities is extremely difficult given the availability of these experts. This issue is universally recognized and has given rise to research experiments and new companies which can use large numbers of lay-persons to perform the scoring and arrive at similar results (Holst et al, 2015). The Crowd Sourced Analysis of Technical Skills (C-SATS) system and company were created specifically to farm out evaluations of videos to large numbers of people who use the Amazon Mechanical Turk platform to perform small jobs for micro-payments. This system can collect evaluations of a simulated or real surgical video from more than 50 evaluators within an hour. The system then applies rules for culling out inappropriate evaluators and computes the average score. Research has shown this average to be statistically equivalent to the scores assigned by a small number of experienced surgeons and surgical educators.

Alternatively, computer simulators are programmed to collect metrics during training exercises and automatically compare these to the measurements of experienced surgeons. However, digital simulator metrics usually measure things that are easy for a computer to measure, but which are not necessarily accurate assessments of the skill levels demonstrated in human performance. Simulator metrics often include “time to complete exercise”, “centimeters of instrument movement”, “duration of excessive force applied”, and “object collision”. All of these are useful indicators of performance, but they do not indicate the degree to which a surgeon has learned to transition from 2D laparoscopic views to 3D robotic views, their proficiency using both hands, and their skill in manipulating instruments (Liu et al, 2015). What simulators measure well is not necessarily what needs to be measured to assess skill levels. Finally, it is very difficult for a simulator to combine all metrics into a single meaningful statement regarding the gestalt of the skill of the student.

Medical and surgical education programs continue to search for assessment methods and tools that will efficiently provide a reliable measure of the performance of the students under training.

ARTIFICIAL INTELLIGENCE TOOLS

The term “artificial intelligence” has been described as a “suitcase word” which contains a wide variety of computational techniques to assist a computer in performing a task that was previously believed to be accessible only to humans (Minsky, 2006). In the decades since it was first coined in 1956 by John McCarthy, AI has been short-hand for techniques like search algorithms, expert systems, rule-based systems, fuzzy logic, constraint satisfaction, planning systems, Bayesian networks, and neural networks (Russell & Norvig, 2015). Each of these was heralded as a revolution that could significantly impact society, commerce, and employment. Each delivered modest advances in its time. But none had the revolutionary effect on society that was portrayed in the news media and science fiction literature.

The latest and most impactful of these techniques is known as deep learning or deep neural networks (LeCun, Bengio & Hinton, 2015). Deep neural networks (DNN) are an extension of the Perceptron network originally posited in 1956 (Rosenblatt, 1956). Modern computer clusters and their availability via cloud services has made it possible to create neural networks with millions of neurons and to make these networks available to any researcher with access to the internet.

The easy availability of these systems raises the question of whether they can be used for performance assessment of surgeons in training. Specifically, could a DNN evaluate skill levels as demonstrated in a video of a specific exercise? Could the DNN be more objective, consistent, and attentive when performing the same assessment as a human evaluator? Would scores provided by a DNN combine the strengths of both subjective human evaluators and objective simulator metrics?

DNN techniques deliver impressive results in recognizing static objects in individual pictures and are beginning to address the recognition of dynamic activities in video. For surgical skills performance, these techniques must (1) recognize stationary objects in the scene, (2) identify the dynamic activities that are demonstrated in the video, and (3) classify the quality of performance of the activity. This paper reports the results of work to accomplish the first of these three objectives.

Neural Network Structures

The field of DNN includes a very large array of different structured networks. Some of the more common are the convolutional neural network (CNN), recurrent neural network (RNN), and long short-term memory neural networks (LSTM). Each of these has been designed for or been found to be effective with a specific type of problem. The CNN and LSTM are both especially powerful in identifying features in pictures and are being applied to the analysis of videos. Since video analysis is a relatively new application of DNN, the research community is still inventing specialized networks for processing these. The services used in this experiment utilized customized versions of the CNN, the general structure of which is shown in Figure 2.

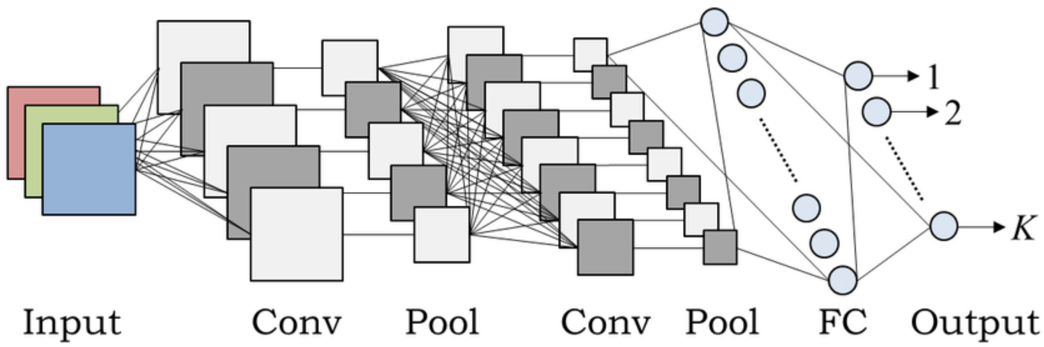


Figure 2. Deep Learning Convolutional Neural Network Architecture

CNNs are characterized by the application of multiple layers of three different categories. These are the convolutional layer, pooling layer, and fully connected layer. In Figure 2 the network possesses the following layers – convolution 1, pooling 1, convolution 2, pooling 2, and fully connected 1. It consists of just five layers with an input layer at the beginning and output layer at the end. Such a network is very useful for illustrating the general structure of a CNN but is capable of processing only the most basic problems. A CNN capable of identifying a large variety of different images would contain many dozens or hundreds of layers.

METHODOLOGY

Programming very specific DNNs for video analysis is a graduate specialization in computer science. This level of expertise is not accessible to most organizations that would benefit from the power of applying these techniques to their problems. Recognizing this situation, several service providers have created online resources that are either pre-built for specific problems or which can be customized with a reasonable amount of programming, usually using online cloud computers, storage, and development tools. For this project we studied the capabilities of several of these services to perform the tasks described above. Specifically, we explored Google Cloud AutoML and Video Intelligence, IBM Watson Visual Recognition, Microsoft Video Indexer, and Amazon Rekognition.

For the first phase of this project we primarily used Google and IBM because of the flexibility of their systems, ease of accessing the tools, and availability of tutorials.

Google's consultants and training programs recommend the levels of effort to be expended in each step when building a DNN solution. Their experience both internally and with external clients has shown that most teams spend far too much time trying to fine-tune the neural network's parameters and far too little collecting data at the beginning. A comparison of these levels of effort as expected by novices and contrasted with expert reality is shown in Figure 3. These steps became the basis of our experimental methodology.

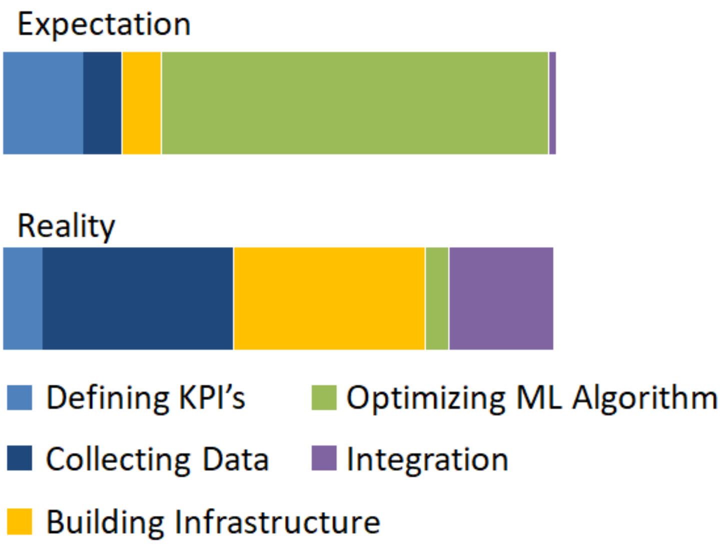


Figure 3. Machine Learning (ML) Effort Allocation (Google, 2018)

Define the KPI's (10%)

Defining the Key Performance Indicators (KPI's) for the network that is to be built is such an obvious step that teams tend to overlook it. For our project, we began with a single KPI, "Can the network correctly score simulation videos?" But after receiving guidance from our Google consultant we devoted some time to discussing what features would lead to a successful DNN product or service. This led us to expand to multiple KPI's that focus on a long-term product life for our work. These include:

- Degree to which the neural network arrives at scores consistent with those of human evaluators.
- Processing rate and cost per video.
- Ease of use of the user interface to access the DNN.
- Long-term availability of the final trained neural network.
- Adaptability of the network to new exercise videos.

Data Collection and Pre-processing (30%)

DNN analysis of videos of surgical training exercises must be based on a collection of existing videos for which objects, activities, and scores have already been applied by a simulator or a human expert as described in an earlier section. The Nicholson Center records videos of exercise performance both with simulators and dry-lab devices on the da Vinci robot. For a previous experiment we recorded over 200 videos of exercises performed on simulators of the da Vinci robot. Each of these were scored by a human expert using the Global Evaluative Assessment of Robotic Skills (GEARS) metric tool. This database of videos and scores form the basis for the training and testing sets for the DNN. Since each video has an associated score from both a computer simulator and a human expert, we intend to explore replicating each of these independently.

The pre-programmed services at each of the providers were all created to solve much simpler problems. As an intermediate step in learning to use these services, we began with the two easier problems with our videos. First, we used the Vision services to identify the objects that appear in the videos. Second, we asked the Video networks to identify the activity that was being performed in the videos. Google Vision & Video Intelligence are existing networks that have been pre-trained to recognize objects from millions of online videos, primarily from YouTube. Therefore, these already recognize tens of thousands of different objects. However, surgical videos contain unique objects that are not usually found in YouTube videos. Therefore, our expectations were that this service might recognize the objects that are being manipulated in the video, but probably not the surgical instruments themselves. The objects that appear in these videos are listed in the first column of Table 1. The exercise name and activity being performed are shown in the second column. The quality metrics on performance from both the simulator and a human evaluator using GEARS are shown in the third column.

Note that a neural network does not really understand the meaning of each of these object names or performance metrics. It is simply creating a mathematical mapping from a set of inputs (video pixels) to a set of outputs (numbers representing an index or a score).

Table 1. Target Metrics for Deep Neural Networks

Objects in Scene:	**Activity Performed:**	**Quality of Performance:**
Surgical Instruments:	*Exercise Name:*	*Simulator Score (numeric):*
Needle Driver	Pick & Place	Time to Complete
Maryland Bipolar	Suture Sponge	Excessive Instrument Force
ProGrasp	Ring & Rail	Instrument Collision
	Ring Transfer	Master Workspace Range
Objects Manipulated:	Matchboard	Instrument Out of View
Needle		Overall Score (weighted sum)
Sponge	*Action Performed:*	
Circular ring	Suturing	*GEARS Score (Likert 1-5):*
Tracking rail	Object Transfer	Depth Perception
Jack (as in a child's game)	Ring on Rail	Bimanual Dexterity
Bowl	Object Positioning	Efficiency
Peg		Force Sensitivity
Letters (A-C)		Robotic Control
Numbers (1-6)		Overall Score (Likert sum)

The videos were collected into a library and clipped to durations which could be processed by the DNN. DNN algorithms analyze videos as a sequence of frames. They are designed to make decisions after a relatively short set of frames, rather than watching extended lengths of video before deciding. The length of segments chosen for this experiment was a combination of the recommendations for the DNN service and our knowledge of the duration of viewing by typical human evaluators.

The videos were not dimensionally trimmed to focus on a specific area or object in the scene. Simulation-based scenes are typically already focused on the objects necessary for the exercise and contain very few extraneous objects that would cause confusion.

Build Infrastructure (35%).

It is common for software developers to create their own local computer systems and networks to host the software that is being developed. This model will not work well for most users of neural networks. The computational capability required to solve most large image and video recognition problems is beyond the resources and technical capabilities of most companies. Installing the necessary software libraries and configuring for their use is also a significant technical challenge. Finally, the amount of storage required for thousands of videos or millions of images can be prohibitive.

When using pre-trained and pre-programmed neural networks, these typically reside on the cloud-based clusters of service providers like Google and IBM. When using these resources, process management software can farm large jobs to thousands of computers, which requires no specific programming by the development team.

Google consultants and training material recommend that a significant amount of time and effort be focused on building the computer infrastructure if this work will be performed on the developers own hardware. But, since we chose to run all this work on the commercial cloud, we simply had to determine the format for data storage and the syntax for accessing the necessary services. Though these have variations across the different providers, the concepts are similar enough that provisioning each becomes progressively easier.

Optimize ML Algorithm (5%)

Machine Learning (ML) is an iterative process. The first pass through any set of data and algorithms will reveal weaknesses and opportunities for improvement. There are several recommended steps that can be taken by customers to improve performance. These include:

- Add more videos of all types.
- Add more videos for the class labels that are performing poorly.
- Adjust labels to a more general class name. (e.g. instrument vs. needle driver)

- Split label classes into distinct, non-overlapping definitions. (e.g. break "bipolar grasper" into bipolar and grasper separately)
- Reduce the number of classes to focus on the most important. (e.g. ignore extraneous objects)
- Identify data leakage. (e.g. videos with text in-scene that gives the name of the instrument)
- Differences in the training, validation, and test sets. (e.g. camera angle, lighting, background in the videos)

Integration into a Solution (20%)

Finally, the investment of time and money into a DNN project should not be undertaken without a plan for integrating the results into a company's products or services. Google defines itself as a machine learning company because they have incorporated over 4,000 ML and DNN algorithms into their various products. They are well on their way to creating a set of products and services that are primarily self-functioning. This will allow them to focus their human resources on pushing the state-of-the-art to invent new technologies and products, as opposed to assigning them to maintaining existing products and services. They counsel all their customers to consider this type of transformation.

A DNN solution for rating surgical simulation videos must have an accessible front-end user interface. It must present itself to those who would use it in a manner that is self-service so the developers do not have to remain directly involved in the daily use of the system. The same applies to the backend hosting of the DNN, it should be reliable, scalable, and require almost no human intervention to keep the system running or to upgrade its hardware and software. Finally, a system needs tools and tutorials so the users of the system can teach themselves to use the DNN. Human-to-human education of the basics should not be necessary. Google has implemented this throughout all their ML and DNN services. Each of them includes manuals, videos, and online courses that can be used to develop mastery.

RESULTS

This research project has three fundamental performance objectives: (1) identify objects, (2) identify actions, and (3) score the quality of performance. This paper reports on the first of these objectives. The pre-trained networks that we used all lack depth in understanding surgical objects. Without retraining of those networks the results are generally poor, though they are all good at recognizing the presence and position of objects in the scene.

Figure 4 illustrates the pre-trained Google Vision API guess at the contents of an exercise called Basic Suture Sponge. The object that we consider a malleable suture sponge has been identified as a box. This is to be expected given its cubic shape and brown coloration. The instrument in view is identified as a weapon, usually associated with a gun or knife. Again, this is a reasonable guess based on the long shape, black and silver coloration, and tapering point at the end. Since the pretrained network is able to identify the location of these objects in the scene, this indicates the feasibility of retraining the network to apply different labels to the objects it can already see. The scene also contains a surgical needle which the network cannot see. It is unlikely that specialized training will lead to the detection of this object.

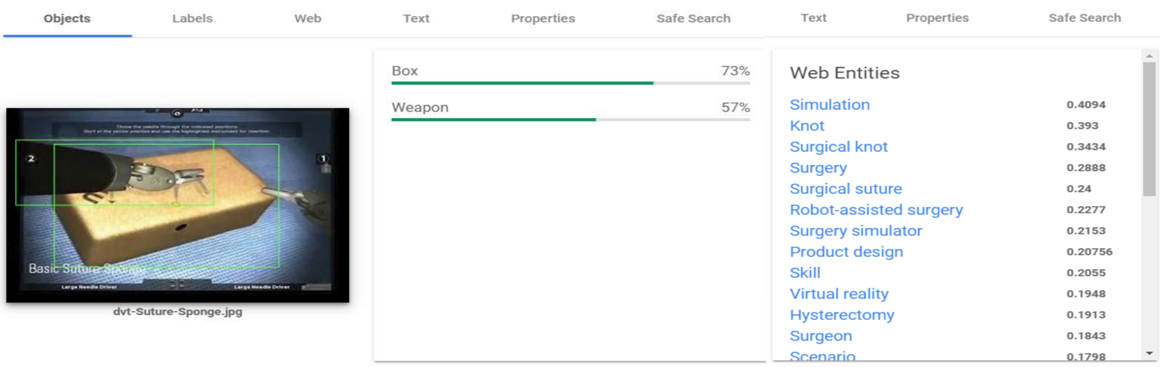


Figure 4. Pre-trained Google Vision API Surgical Object Recognition.

Figure 4 also illustrates a feature that is only available from the Google service. On the right side, it has searched the internet for images like the one provided, then extracted the descriptors attached to those pictures and returned a list. These labels themselves are not trained into the DNN, they are image meta data applied by humans on an internet web page. In this case, these labels are very accurate descriptions of the objects themselves or the activities they would be used for.

Figure 5 provides the results from each of the service providers in processing the same image. Each service's understanding of the objects in the scene is different, from padlock to weapon to cutlery to vise. It is clear that prior to specialized re-training, none of them are familiar with surgical objects.

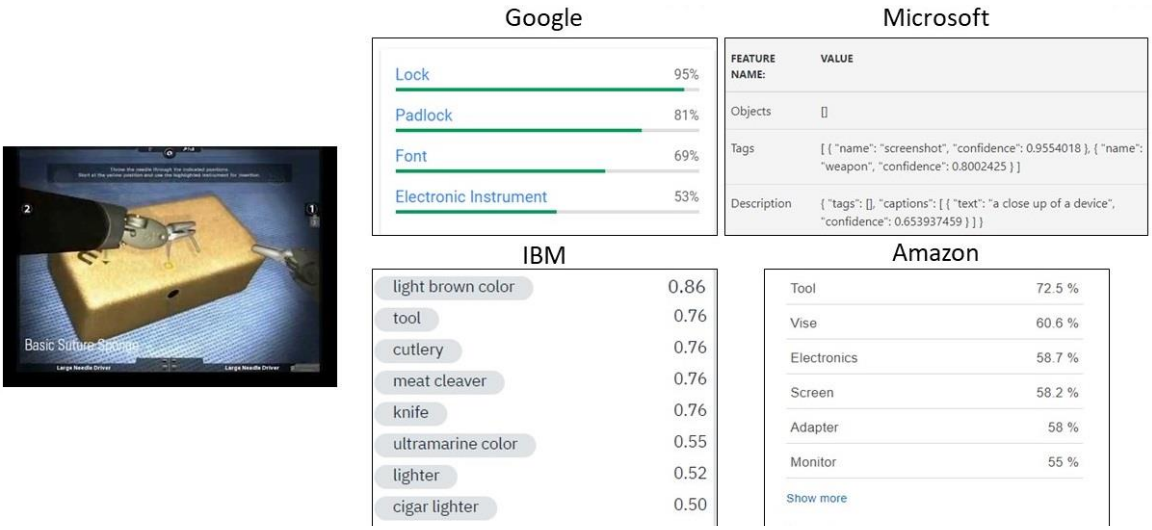
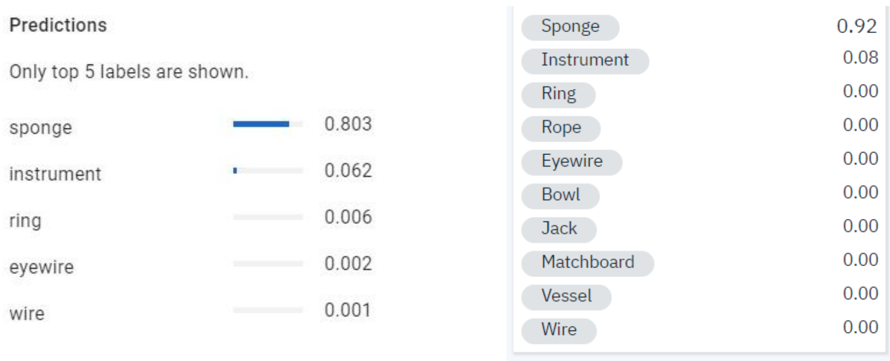


Figure 5. Pre-trained, Pre-programmed DNN Services Results.

Using this information as a starting position, we collected a large set of images from multiple exercises in the surgical simulator (420 images, containing 11 objects, appearing in 10 different exercises). These were labeled with the objects in the scene and used as the training set for both the Google and IBM services. The DNN structure provided by these services were not changed, only the decision weights were re-trained. After being re-trained on this image set, the two networks were then able to apply much more relevant labels to the images, as shown in Figure 6. Microsoft and Amazon services cannot be custom trained through a web interface. Both require users to write a custom application that calls their DNN APIs, which we did not do in the first phase of this study.



(a) Google (b) IBM

Figure 6. Re-trained Google and IBM Vision API Surgical Object Recognition.

DISCUSSION

As offered, without modification, these services are capable of accurately identifying objects that are commonly found in online videos and photos. They can detect the presence of objects in surgical simulation videos but are not able to correctly identify those objects. However, these DNN services are so robust that they can be easily trained to identify a unique set of objects that they had not previously been exposed to. Google and IBM services can be retrained using a web interface requiring no programming. These providers have been successful at creating a robust and flexible service that can put the most complex functionality of DNNs within reach of businesses that can use them for specialized applications.

The services used in this study (Google, IBM, Microsoft, and Amazon) all offer a private, cloud-based, virtual machine in their own computer centers. These execution environments are sufficiently secure for most commercial applications, to include the evaluation of surgical training videos. However, to be used for military problems, these providers would have to create computer centers and connections that comply with DOD information security standards. Therefore, military and intelligence applications of these tools may lag applications by commercial and research customers.

This paper reports the results of the first phase of this research project, identifying and correctly labeling objects in surgical simulation videos. Accomplishing the second phase of recognizing the activity in the video should be possible with these services as well. However, the third phase, scoring the quality of the action being performed, is an active research area and is unlikely to be achievable with these services. It will require creating a customized network architecture and training protocol, which we are currently pursuing with a university research team.

CONCLUSIONS

Creating effective deep neural networks is a task that requires specialized education and experience, huge volumes of training data, extensive computer infrastructure, and years of experimentation. As such, these cannot be constructed or trained from scratch by most business organizations. Pre-programmed and pre-trained networks are offered from multiple commercial providers. These can be used as-is or can be modified to address specific problems and specific data domains, such as the surgical simulation data described here.

Though there are thousands of potential government and business applications of neural networks, these will rarely be satisfied by a company creating their own networks from scratch. Specialized providers like those used in this study will emerge as the source of these tools, following a pattern similar to the commercial availability of word processors (e.g. Microsoft) and customer relationship management (e.g. Salesforce) software. Word processors and CRM systems began as locally installed applications and are now migrating to cloud-based services. Deep neural network applications are beginning as cloud-based applications because of the extreme compute and storage requirements. Very few users of these tools will have the resources to host them completely in-house on privately owned infrastructure.

We are satisfied with the accessibility and security of these commercial services for surgical training videos. We also find it possible to modify these services with a reasonable amount of effort and specialized expertise. The results of this research experiment are on a path to becoming a customized service that we can use in our business or even expose as a commercial service for other institutions with similar needs.

REFERENCES

Abboudi, H., Khan, M., Aboumarzouk, O., Guru, K., Challacombe, B., Dasgupta, P., & Ahmed, K. (Feb 2013). “Current Status of Validation for Robotic Surgery Simulators – a Systematic Review.” *BJU International*, 111(2).

Davis, D., O’Brien, M., Freemantle, N., Wolf, F., Mazmanian, P., & Taylor-Vaisey, A. (1999). “Impact of Formal Continuing Medical Education”, *Journal of the American Medical Association*, 282(9), 867-874.

Davis, D., Mazmanian, P., Fordis, M., Van Harrison, R., Thorpe, K., & Perrier, L. (2006). “Accuracy of Physician Self-assessment Compared with Observed Measures of Competence: A Systematic Review”, *Journal of the American Medical Association*, 296, 1094-1102.

Goh, A., Goldfarb, D., Sander, J., & Dunkin, B. (Jan 2012). “Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills”. *Journal of Urology*, 187(1), 247-252.

Google Cloud Team. (2018). “Machine Learning with TensorFlow on Google Cloud Platform”, online training course. Available at: https://www.coursera.org/specializations/machine-learning-tensorflow-gcp.

Holst, D., Kowalewski, T., White, L., Brand, T., Harper, J., Sorensen, M., Truong, M., Simpson, K., Tanaka, A., Smith, R., & Lendvay, T. (2015). “Crowd-Sourced Assessment of Technical Skills: Differentiating Animate Surgical Skills Through the Wisdom of Crowds”. *Journal of Endourology*, 29(40).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). “Deep Learning”. *Nature*, 521(7553), 436-445.

Liu, M. & Curet, M. (2015). “A Review of training research and virtual reality simulators for the da Vinci surgical system”. *Teaching and Learning in Medicine*, 27(1).

Martin J.A., Regehr G., & Reznick R. (1997). “Objective structured assessment of technical skill (OSATS) for surgical residents.” *British Journal of Surgery*, 84(2), 273–278.

Minsky, M. (2005). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster.

Russell, S. and Norvig, P. (2015). *Artificial Intelligence: A Modern Approach*. Noida, India: Pearson Publishing.

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". *Psychological Review*, 65(6), 386-408.

Tch, A. (Aug 4, 2017). “The Mostly Complete Chart of Neural Networks, Explained.” *Towards Data Science*. Available online at: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464.

Tian, J., Atkinson, N., Portnoy, B., & Gold, R.. (2007). “A Systematic Review of Evaluation in Formal Continuing Medical Education.” *Journal of Continuing Medical Education in the Health Professions*, 27(1), 16-27.

van Hove P.D., Tuijthof G.J., Verdaasdonk E.G., Stassen L.P., & Dankelman J. (2010). “Objective assessment of technical surgical skills.” *British Journal of Surgery*, 97(7), 972–987.

CLOUD-BASED NEURAL NETWORK RESOURCES

Google Video Intelligence. https://cloud.google.com/video-intelligence/

IBM Watson Visual Recognition. https://www.ibm.com/watson/services/visual-recognition/

Microsoft Video Indexer. https://azure.microsoft.com/en-us/services/media-services/video-indexer/

Amazon Rekognition. https://aws.amazon.com/rekognition/

Google Machine Learning Courses. https://www.coursera.org/specializations/machine-learning-tensorflow-gcp

IBM Data Science Courses. https://www.coursera.org/specializations/ibm-data-science-procfessional-certificate