COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
2021, VOL. 9, NO. 3, 313–321
https://doi.org/10.1080/21681163.2020.1835548

The AI-Medic: an artificial intelligent mentor for trauma surgery

Edgar Rojas-Muñoz[a], Kyle Couperus[b] and Juan P. Wachs[a]

[a]School of Industrial Engineering, Purdue University, West Lafayette, IN, USA; [b]Department of Emergency Medicine, Madigan Army Medical Center, Tacoma, WA, USA

ABSTRACT

Telementoring generalist surgeons as they treat patients can be essential when in situ expertise is not available. However, unreliable network conditions, poor infrastructure, and lack of remote mentors availability can significantly hinder remote intervention. To guide medical practitioners when mentors are unavailable, we present the AI-Medic, the initial steps towards an intelligent artificial system for autonomous medical mentoring. A Deep Learning model is used to predict medical instructions from images of surgical procedures. An encoder-decoder model was trained to predict medical instructions given a view of a surgery. The training was done using the Dataset for AI Surgical Instruction (DAISI), a dataset including images and instructions providing step-by-step demonstrations of 290 different surgical procedures from 20 medical disciplines. The predicted instructions were evaluated using cumulative BLEU scores and input from expert physicians. The evaluation was performed under two settings: with and without providing the model with prior information from test set procedures. According to the BLEU scores, the predicted and ground truth instructions were as high as 86 ± 1% similar. Additionally, expert physicians subjectively assessed the algorithm subjetively and considered that the predicted descriptions were related to the images. This work provides a baseline for AI algorithms assisting in autonomous medical mentoring.

ARTICLE HISTORY
Received 11 September 2020
Accepted 7 October 2020

KEYWORDS
Datasets; neural networks; telementoring; surgery

1. Introduction

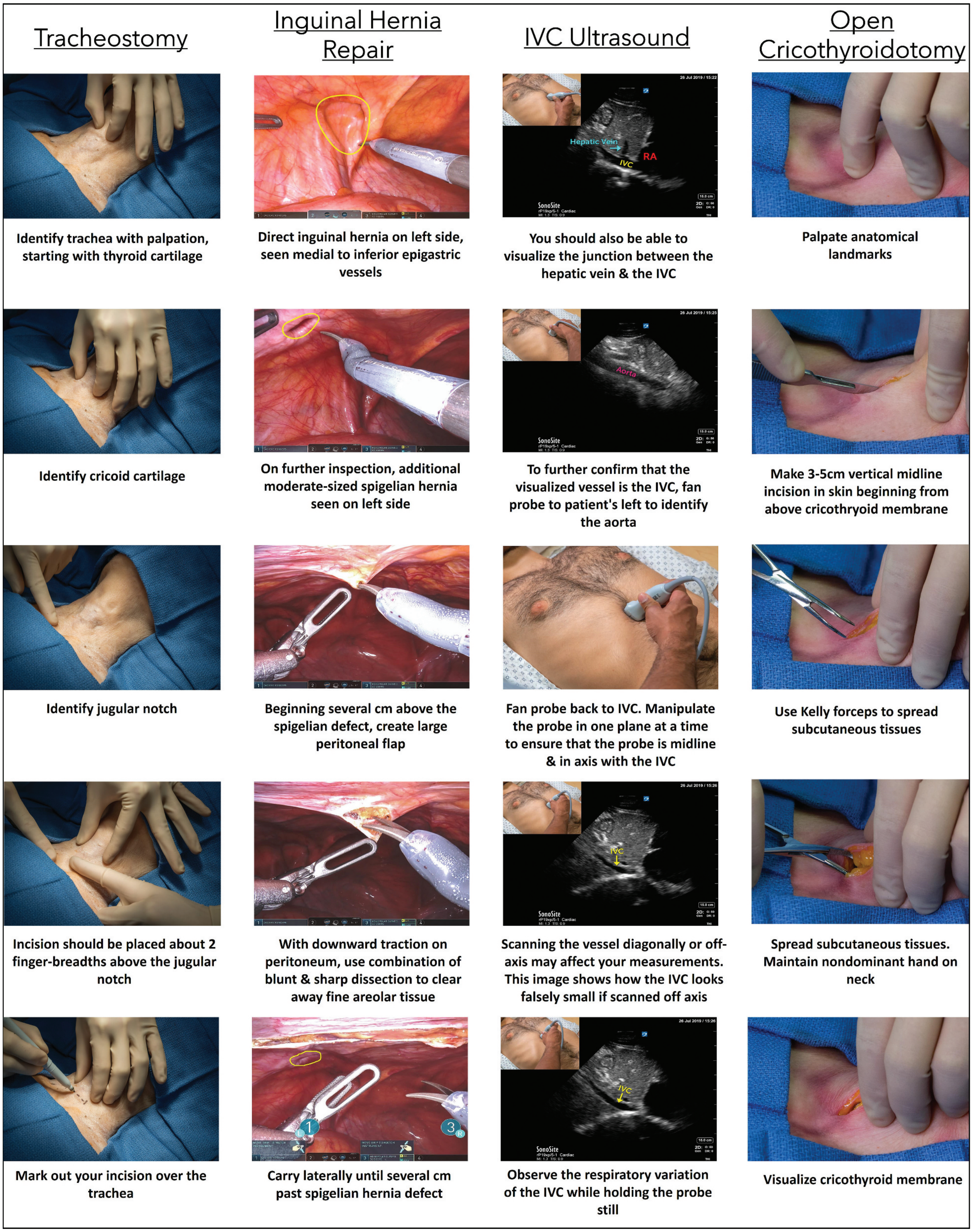
In telementoring, a surgeon performing an operation receives guidance remotely from an expert using telecommunications. Such technology can support life saving interventions in rural, remote and even austere settings (Sebajang et al. 2006; Greenberg et al. 2015). Techniques, such as telementoring through augmented reality and speech have been explored to provide generalist surgeons with remote supervision when no expert specialist is available on-site (Rojas-Muñoz et al. 2020). Two aspects are fundamental requirements for these techniques: the availability of the expert that provides medical guidance, and having a reliable communication medium. Nonetheless, such requirements cannot always be satisfied (e.g. unreliable network conditions, cyber-attacks, etc). When these requirements cannot be fulfilled (Bilgic et al. 2017), a fallback artificial intelligence (AI) mechanism could provide medical practitioners with the necessary hands-on knowledge to complete the surgical procedure. However, the development of AI-based autonomous mentoring frameworks applied to medicine has been limited due to the lack of robust predicting models and quality of datasets that are required to train such models. Such datasets can include step-by-step demonstrations for completing surgical procedures from various medical specialities.

This work introduces the AI-Medic, an AI system for autonomous medical mentoring. The system uses a Deep Learning (DL) model to predict medical instructions from images of surgical procedures. The model was training using the Dataset for AI Surgical Instruction (DAISI; https://engineering.purdue.edu/starproj/_daisi), a dataset that includes images and instructions providing step-by-step demonstrations of surgical procedures. DAISI includes images and text descriptions of procedures from 20 medical disciplines. Each image-text pair describes the surgical manoeuvres or required instruments to complete one step in the procedure. Thedataset was created based on feedback from 20 expert physicians from various medical centres, extracting data from academic medical textbooks related to the surgical technique, and acquiring imagery manually. Such a dataset addresses the shortage of datasets for autonomous medical mentoring. Figure 1 showcases images from four procedures in DAISI.

Our AI-Medic uses the DAISI dataset to train an encoder-decoder DL architecture capable of predicting instructions given the current view of a surgery. The model takes images of medical procedures as input, and outputs text descriptions with the instructions to perform next by a surgeon. The encoder-decoder approach uses a ConvNet as the encoder network, and a Recursive Neural Network (RNN) as the decored network. The ConvNet extracts and encodes visual features from the input images, and the RNN decodes these visual features into text descriptions. The instructions predicted by the AI-Medic were evaluated using Bilingual Evaluation Understudy (BLEU) scores (Papineni et al. 2002) and subjective input from expert physicians. The results presented in this work serve as a baseline for future AI algorithms that can be used as surrogate human mentors.

The paper proceeds as follows: Section 2 reviews approaches for medical autonomous guidance and related datasets. Section 3 describes DAISI and how it was used to train the AI-Medic's DL model. Section 4 presents and discusses the results obtained from evaluating the AI-Medic. Finally, Section 5 concludes the paper.

CONTACT Edgar Rojas-Muñoz emuoz@purdue.edu School of Industrial Engineering, Purdue University, West Lafayette, IN, USA



Tracheostomy

Identify trachea with palpation, starting with thyroid cartilage

Identify cricoid cartilage

Identify jugular notch

Incision should be placed about 2 finger-breadths above the jugular notch

Mark out your incision over the trachea

Inguinal Hernia Repair

Direct inguinal hernia on left side, seen medial to inferior epigastric vessels

On further inspection, additional moderate-sized spigelian hernia seen on left side

Beginning several cm above the spigelian defect, create large peritoneal flap

With downward traction on peritoneum, use combination of blunt & sharp dissection to clear away fine areolar tissue

Carry laterally until several cm past spigelian hernia defect

IVC Ultrasound

You should also be able to visualize the junction between the hepatic vein & the IVC

To further confirm that the visualized vessel is the IVC, fan probe to patient's left to identify the aorta

Fan probe back to IVC. Manipulate the probe in one plane at a time to ensure that the probe is midline & in axis with the IVC

Scanning the vessel diagonally or off-axis may affect your measurements. This image shows how the IVC looks falsely small if scanned off axis

Observe the respiratory variation of the IVC while holding the probe still

Open Cricothyroidotomy

Palpate anatomical landmarks

Make 3-5cm vertical midline incision in skin beginning from above cricothryoid membrane

Use Kelly forceps to spread subcutaneous tissues

Spread subcutaneous tissues. Maintain nondominant hand on neck

Visualize cricothyroid membrane

Figure 1. Example of five different images and their associated textual descriptions from four different procedures in the DAISI dataset. The dataset includes images from 20 disciplines such as emergency medicine, and ultrasound-guided diagnosis.

2. Background

Telementoring systems can be used to deliver expert assistance remotely (Kotwal et al. 2016). Several studies demonstrate that such systems can provide surgeons with specialised assistance in austere settings when no expert is on-site (Sebajang et al. 2005, 2006). However, telementoring platforms rely on remote specialists being available to assist, which often is not possible

(Geng et al. 2019). A possible approach to convey guidance when no mentor is available is by incorporating AI into telementoring systems (de Araújo Novaes and Basu 2020). In this way, a virtual intelligent surrogate mentor can be activated to guide health practitioners when the mentor is not available.

AI algorithms have been previously explored as means to assist users during complex and time sensitive decision-making procedures (Cortés et al. 2007). In the healthcare domain, for example, AI has been typically used in the diagnosis and prognosis of diseases (Patel et al. 2020; Mishra and Banerjee 2020). Typically, the diagnosis is given via predictions of an AI model, trained using a dataset of medical records (e.g. radiology images, metabolic profiles). However, recent approaches have also incorporated AI into surgical instruction. The Virtual Operative Assistant is an example of an automated educational feedback platform (Mirchi et al. 2020). This platform was developed to provide automated feedback to neurosurgeons performing a virtual reality brain tumour resection task. Other examples include AI to analyse surgical performance during virtual reality spine surgeries (Bissonnette et al. 2019), and integration with augmented reality to provide surgical navigation during surgery (Auloge et al. 2019; Jha and MB 2019).

Our work follows an approach in which AI is used to train models capable of predicting medical image descriptions (Kisilev et al. 2016; Singh et al. 2019; Alsharid et al. 2019) Similar examples include the ImageCLEFcaption challenge, which focuses on using AI to obtain text descriptions from radiology images (Lyndon et al. 2017; Pelka et al. 2019; Xu et al. 2019). Likewise, IU X-RAY and PEIR GROSS are examples of public datasets for radiology image captioning (Jing et al. 2017; Pavlopoulos et al. 2019). These algorithms, however, are trained to generate captions that describe the content of the images using techniques such as template-based image captioning, retrieval-based image captioning, and novel caption generation (Hossain et al. 2019). Instead, our work generates captions representing instructions in a task by learning visual-semantic correspondences between images and instructions using an architecture similar to the one created by Karpathy and Fei-Fei (2015).

3. Methods

The methodology to create the AI-Medic includes: 1) the creation of a curated dataset of medical images and their respective step-by-step descriptions; and 2) the use of such a dataset to train a DL framework that generates medical instructions from images.

3.1. Creating a dataset for AI surgical instruction

The DAISI dataset contains 17,339 colour images and text descriptions of instructions to perform surgical procedures. DAISI contains one example for each of the 290 medical procedures from 20 medical disciplines including ultrasound-guided diagnosis, trauma and gynaecology. The image-text pairs from the dataset were compiled from: (a) medical images and instructions from the Thumbroll app (thumbroll LLC 2020), a medical training app designed by physicians from Washington University School of Medicine, Stanford Health Care; John Hopkins University, UCLA, and University of Southern California. Overall, we acquired 17,138 images with descriptions from various medical specialities (e.g. General Surgery, Internal Medicine), levels of medical training (e.g. clinical Medical Doctor trainee, senior Resident), and medical occupations (e.g. occupational therapy, osteopathic medicine). (b) Afterwards, we extracted 125 images and captions from anatomy textbooks using *PDFFigCapX* (Li et al. 2019); (c) Lastly, we used a patient simulator (Tactical Casualty Care Simulator 1, Operative Experience) to acquire an additional set of 76 images from procedures as chest needle decompression and intraosseous needle placement. The dataset has 1086 duplicate instructions (i.e. the same instruction from more than one image), which represent approximately 6% of the instructions in the dataset. Additionally, no restriction was imposed regarding the length of the text descriptions.

3.2. Training an intelligent agent for autonomous mentoring

We used DAISI to train a DL model for autonomous mentoring. The algorithm receives images from medical procedures as input, and predicts an instruction associated with it. To generate text information from images, an encoder-decoder DL approach using a ConvNet and a Recursive Neural Network (RNN) was adopted. The ConvNet extracts and encodes visual features from the input images, and the RNN decodes these visual features into text descriptions (see Figure 2).

Captioning techniques require a 1 vocabulary containing the words appearing in the dataset at least *N* times (defined by the *Word Count* parameter). This constrains the words used to generate the instruction to a fix set. A token character (UNK) replaced all words in the training set that appeared fewer times than the specified *Word Count* value. Our encoder-decoder architecture is based on NeuralTalk2 (Karpathy and Fei-Fei 2015). We use the VGG16 model as the encoder network (Simonyan and Zisserman 2014). This model includes 13 convolutional layers with 5 pooling layers in-between. The convolutional layers use 3 × 3 convolutional filters to locate interest features in the images, and the pooling layers reduce these features' dimensionality. All hidden layers are equipped with Rectified Linear Units (ReLU). We performed cross validation using the Adam adaptive learning rate optimisation to find individual learning rates for each parameter in the ConvNet (Kingma and Ba 2014). Finally, 4 fully connected layers are used to describe each image with a 1000-dimensional latent vector representation. We then use a Bidirectional Recurrent Neural Network (BRNN) as the decoder network to generate the text instructions (Schuster and Paliwal 1997). The BRNN predicts instructions not only by receiving the ConvNet's final latent vector, but also by leveraging context around the word. This context is determined via forward and backward hidden states (h_t^f and h_t^b, respectively) at each index t ($t = 1\ T$), which denotes the position of a word in a sentence. Therefore, the BRNN predicts semantically correct sentences based on the ConvNet's latent vector and the current word's context. This step is necessary to create semantically coherent sentences rather than disconnected words describing different aspects of the image. The BRNN's 1 formulation follows:

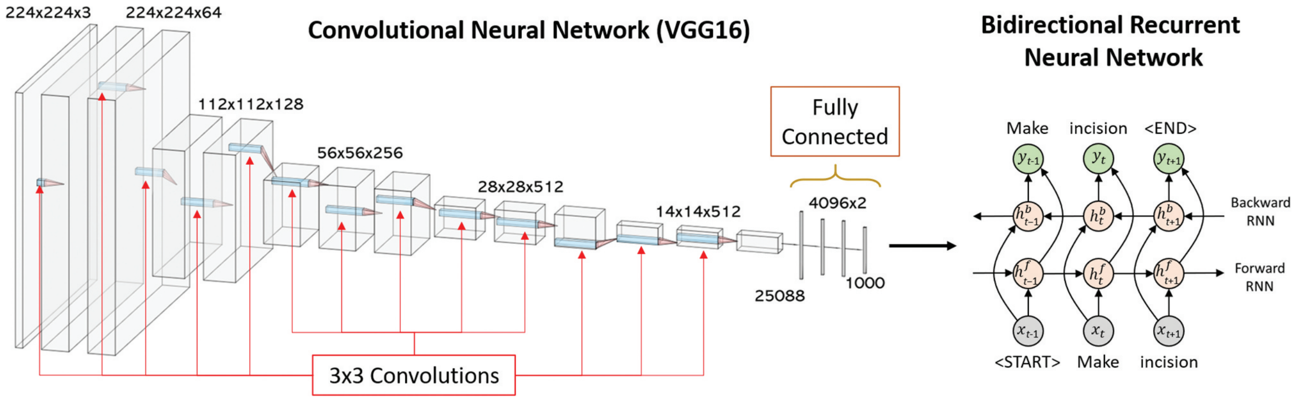


Figure 2. Schematic of our encoder-decoder architecture. The ConvNet obtains vectors representing input images. These vectors are then used in the training of a BRNN that learns to predict surgical instructions. The VGG16 schematic was generated via LeNail's NN-SVG (2019).

$$c_v = W_{hi}[ConvNet_\theta(Img)] \tag{1}$$

$$h^f_t = ReLU(W_{hx}x_t + W_{hf}h^f_{t-1} + b_f + 1(iter = 1) \bullet b_v) \tag{2}$$

$$h^b_t = ReLU(W_{hx}x_t + W_{hb}h^b_{t+1} + b_b + 1(iter = 1) \bullet b_v) \tag{3}$$

$$y_t = Softmax(W_{ho}(h^f_t + h^b_t) + b_o) \tag{4}$$

Where, W_{hi}, W_{hx}, W_{hf}, W_{hb}, W_{ho}; b_f, b_b, and b_o are the parameters and biases to be learned by the model. $ConvNet_\theta(Img)$ is the ConvNet's final latent vector of the image *Img*. Thus, the image context vector c_v provides the BRNN with information from the input image. This context vector c_v is provided only during the first iteration ($iter = 1$), as suggested by Karpathy and Fei-Fei (2015). The x_t and y_t vectors contain probabilities of each word in the vocabulary to be the word at the index *t*. The output vector y_t is used as x_{t+1} in the next iteration. In the first iteration, the output vector y_t depends only on the context vector c_v, as x_t takes a special initialisation value (START) and h^f_t and h^b_t are initialised to 0. This formulation allows the model to predict more than one candidate instruction per image. The probability of each candidate being the correct instruction decreases for each additional prediction.

3.3. Evaluating the artificial intelligent mentor

We evaluate our AI-Medick and provide a benchmark for future AI surgical mentors. We validated our approach using four test folds. For each of these folds, we randomly divided the 290 procedures into training and testing sets based on their number of images: approximately 10% of the images of the entire dataset were separated to be used as test set. Additionally, we conducted *Inter-procedure* and *Intra-procedure* evaluations. For the *Inter-procedure* setting, the model had no prior information regarding the procedures in the test set. For the *Intra-procedure* setting, a fraction of the images *P* in the same procedure were assigned to the training set, while the rest remained in the test set. The test set consisted of every $\frac{1}{P}$ images from each procedure. In our case, *P* was set to 0.5. While the *Intra-procedure* setting reduced generalisability among procedures, it enhanced performance for procedures in the test set. Table 1 presents the distribution of image-text pairs into training and test sets, as well as the size of resulting vocabulary for each fold. Finally, we evaluated our AI mentor using three combinations of the *Word Count* parameter: 3, 5, and 7. Table 2 showcases the size of the vocabulary constructed for each fold, for the respective *Word Count* value.

To evaluate the AI-Medic's performance, the BLEU metric was computed between the predicted and the ground truth instructions. This is a state-of-the-art metric to evaluate text production models related to image captioning and machine translation (Papineni et al. 2002). BLEU computes a 1-to-100 similarity score by comparing two sentences (s_1 and s_2, candidate and reference respectively) at the word *n*-gram level, i.e. analysing contiguous sequences of *n* words in a text. For instance, s_1 and s_2 will have perfect BLEU 1-gram score if all the words from s_1 appear in s_2. Similarly, they will have a perfect BLEU 2-gram score if all possible combinations of two words from s_1 appear in s_2. Contrarily, the BLEU score between s_1 and s_2 will decrease for each *n*-gram that is in s_2 but is not in s_1. We report cumulative BLEU scores for 1-grams to 4-grams for the model's top five candidate predictions, as they have reported correlations with human judgements (Ward and Reeder 2002).

Table 1. Distribution of image-text pairs into training and test sets per testing fold and testing approach (*inter-procedure* and *intra-procedure*).

	Number of images			
	Inter-procedure		Intra-procedure	
Fold Number	Training	Testing	Training	Testing
F1	13,232	1354	13,909	677
F2	13,302	1284	13,944	642
F3	13,256	1330	13,921	665
F4	13,284	1302	13,935	651

Table 2. Vocabulary size for three different *word count* values.

Fold Number	3	5	7
F1	2217	2214	2208
F2	2231	2226	2219
F3	2215	2213	2209
F4	2219	2211	2207

Finally, expert emergency physicians evaluated the algorithm's performance subjectively. The experts were selected randomly from a pool of emergency physicians working in an army medical centre (Madigan Army Medical Center). We randomly selected 16 images from emergency medicine procedures in the test set and their predicted instructions (e.g. Chest Tube Placement, Cardiopulmonary Resuscitation). Afterwards, we used a survey to rate how related was each image to its predicted instruction. Each question in the survey included an image from a medical procedure, the name of the procedure; the instruction predicted by the AI-Medic, and five options rating the relation between the predicted description and the image. The physicians ranked this relation using a normalised scale: 'Very Related' = 1, 'Related' = 0.75, 'Somewhat Related' = 0.5; 'Not Related' = 0.25, and 'Impossible to Tell' = 0. In addition, the physicians were asked to provide their own instructions for the steps depicted in the images. This was done to analyse how consistent the physicians were on describing the step to perform.

4. Results & discussion

Figure 3 shows the resulting instructions predicted by the AI-Medic. The predicted instruction is written inside the images, whereas the ground truth instruction is written below. The predicted instructions were semantically correct because of the relations between images and captions created in the network's embedding space.

Figure 4 reports the cumulative BLEU scores for *Inter-procedure* and *Intra-procedure* testing. The captions predicted by our model obtained up to 86 ± 1% 1-gram and 36 ± 1% 4-gram BLEU scores. Our results surpassed those reported in state-of-the-art approaches for medical instructions prediction (Lyndon et al. 2017). Overall, the BLEU scores were slightly lower for lower *Word Count* values. A potential reason is that an increased-size vocabulary reduced the chances of learning meaningful relations between the images and the text descriptions. Our algorithm tackles a challenging problem due to the interclass variance among different medical procedures, which

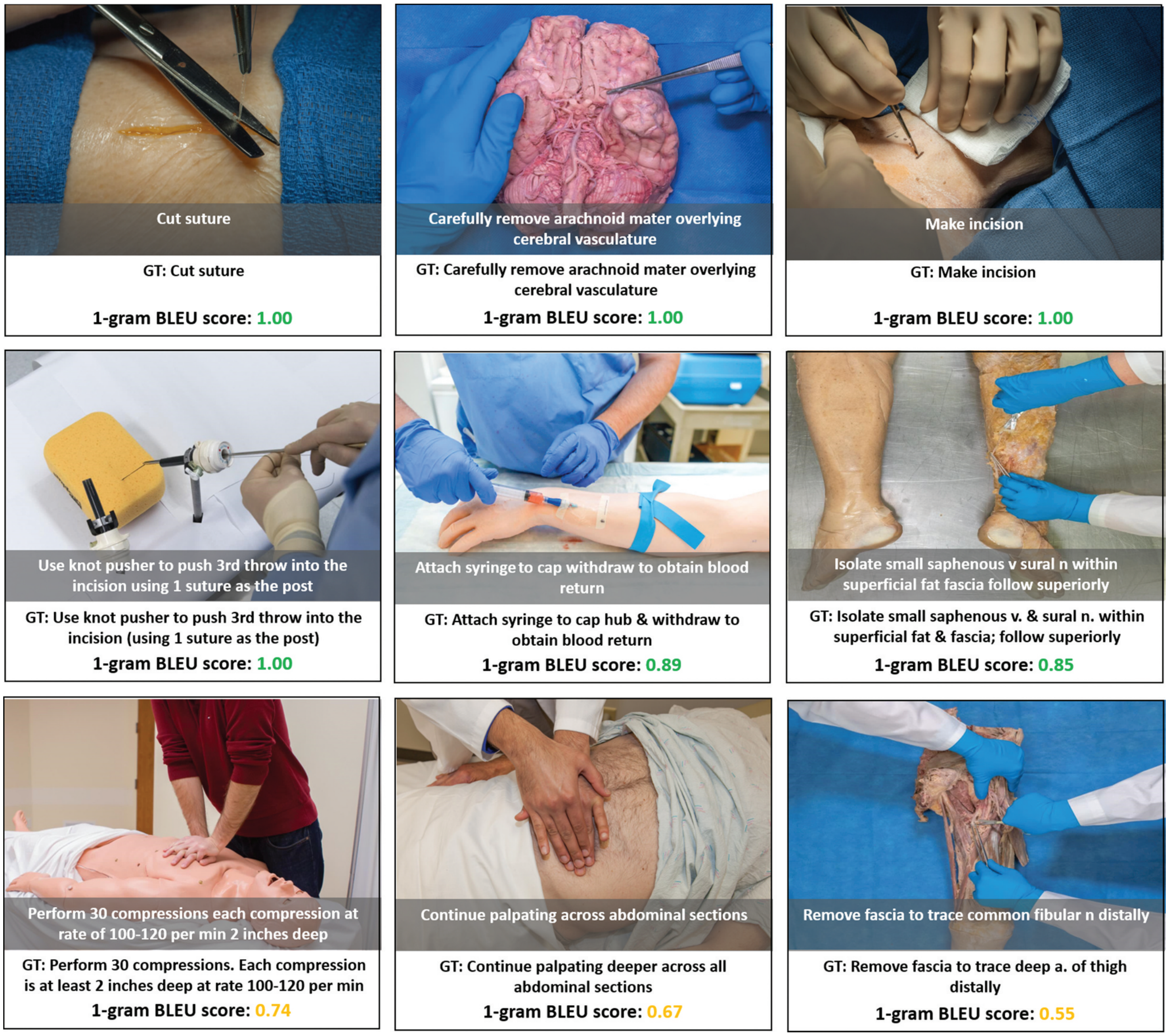

Cut suture
GT: Cut suture
1-gram BLEU score: 1.00
Carefully remove arachnoid mater overlying cerebral vasculature
GT: Carefully remove arachnoid mater overlying cerebral vasculature
1-gram BLEU score: 1.00
Make incision
GT: Make incision
1-gram BLEU score: 1.00
Use knot pusher to push 3rd throw into the incision using 1 suture as the post
GT: Use knot pusher to push 3rd throw into the incision (using 1 suture as the post)
1-gram BLEU score: 1.00
Attach syringe to cap withdraw to obtain blood return
GT: Attach syringe to cap hub & withdraw to obtain blood return
1-gram BLEU score: 0.89
Isolate small saphenous v sural n within superficial fat fascia follow superiorly
GT: Isolate small saphenous v. & sural n. within superficial fat & fascia; follow superiorly
1-gram BLEU score: 0.85
Perform 30 compressions each compression at rate of 100-120 per min 2 inches deep
GT: Perform 30 compressions. Each compression is at least 2 inches deep at rate 100-120 per min
1-gram BLEU score: 0.74
Continue palpating across abdominal sections
GT: Continue palpating deeper across all abdominal sections
1-gram BLEU score: 0.67
Remove fascia to trace common fibular n distally
GT: Remove fascia to trace deep a. of thigh distally
1-gram BLEU score: 0.55

Figure 3. Examples of instructions predicted by the AI-Medic. The predicted instruction is in white font, inside the images. The ground truth (GT) instruction is written below. The approach calculates the BLEU scores after removing special characters (e.g. punctuation marks). High and average scores are highlighted in green and yellow, respectively.

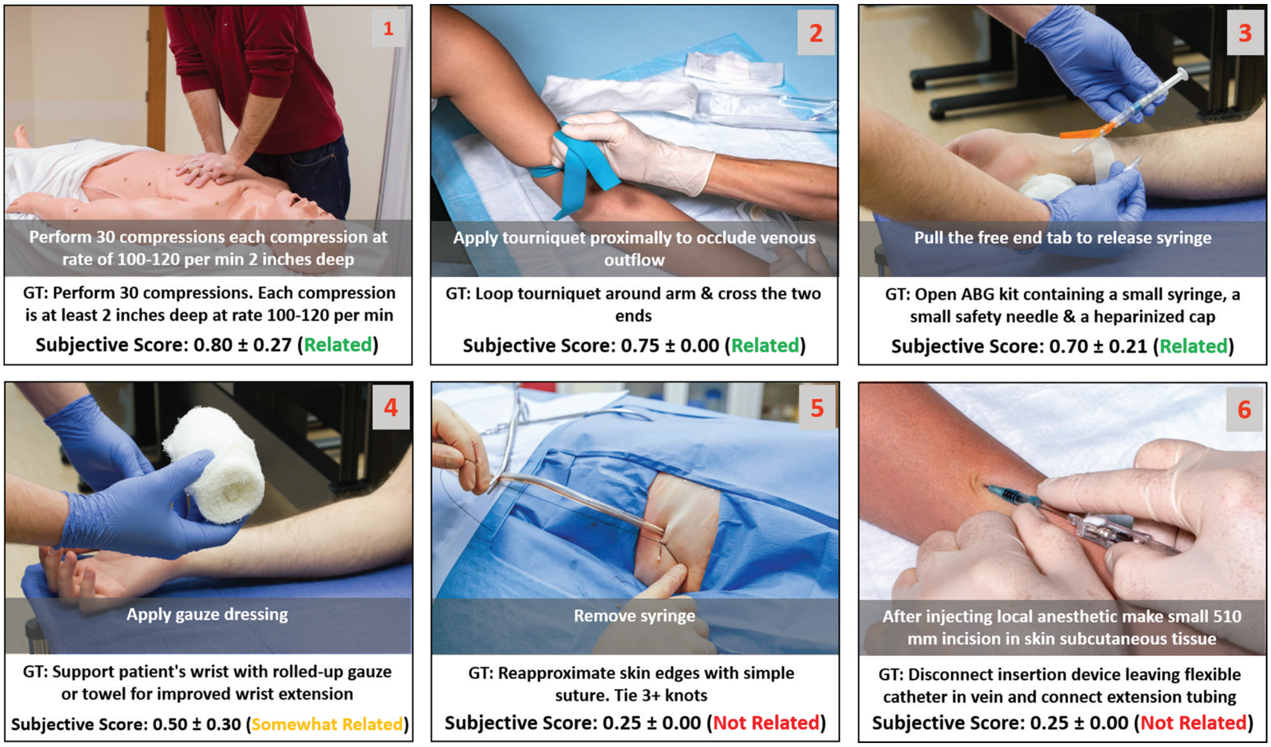

1
Perform 30 compressions each compression at rate of 100-120 per min 2 inches deep
GT: Perform 30 compressions. Each compression is at least 2 inches deep at rate 100-120 per min
Subjective Score: 0.80 ± 0.27 (Related)
2
Apply tourniquet proximally to occlude venous outflow
GT: Loop tourniquet around arm & cross the two ends
Subjective Score: 0.75 ± 0.00 (Related)
3
Pull the free end tab to release syringe
GT: Open ABG kit containing a small syringe, a small safety needle & a heparinized cap
Subjective Score: 0.70 ± 0.21 (Related)
4
Apply gauze dressing
GT: Support patient's wrist with rolled-up gauze or towel for improved wrist extension
Subjective Score: 0.50 ± 0.30 (Somewhat Related)
5
Remove syringe
GT: Reapproximate skin edges with simple suture. Tie 3+ knots
Subjective Score: 0.25 ± 0.00 (Not Related)
6
After injecting local anesthetic make small 510 mm incision in skin subcutaneous tissue
GT: Disconnect insertion device leaving flexible catheter in vein and connect extension tubing
Subjective Score: 0.25 ± 0.00 (Not Related)

Figure 4. Cumulative *n*-gram BLEU scores. Our model was evaluated using three *word count* values (3, 5, 7) and two testing approaches (*Inter-procedure, Intra-procedure*). The model obtained up to 86 ± 1% 1-gram, and up to 36 ± 1% 4-gram BLEU scores.

in turns has an impact the prediction capability of the network. As a reference value, the BLEU 1-gram score when comparing the ground truth instructions with descriptions constructed using random words from the vocabulary is less than 0.1%. Therefore, our results show an improvement over random guess.

Five expert physicians completed our subjective evaluation, for a total 80 responses. The physicians reported having 11.2 ± 3.3 years of medical expertise. On average, the physicians considered the predicted instructions to be ‘Somewhat Related’ to the medical images (0.51 ± 0.32). While this is an encouraging result, a drastic improvement is still required for useful AI mentoring for surgery. Their evaluations followed three main trends. The first trend are descriptions considered as correct predictions: they were similar to the ground truth and physicians considered them as adequate guidance (e.g. Figure 5, examples 1, 2, and 3). The second trend were descriptions that were not similar to the ground truth, but were consider as adequate guidance by the physicians. These descriptions included key elements from the image (e.g. gauze in Figure 5 example 4), but did not use the phrasing of the ground truth. The third trend comprehends

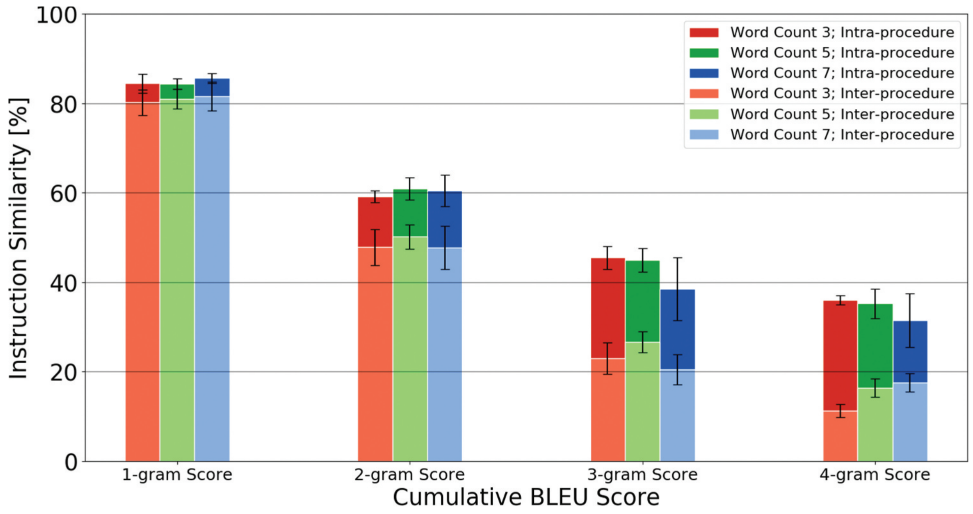


Figure 5. Subjective evaluation by physicians of the instructions predicted by the AI-Medic. High, average, and low scores are highlighted in green, yellow and red font, respectively.

descriptions considered as incorrect predictions (e.g. Figure 5, examples 5 and 6).

Nonetheless, similar discrepancies were found when analysing the instructions provided by the expert physicians. For instance, once physician provided the same instruction as the one predicted by the AI-Medic in Figure 5 example 4 (‘*Apply gauze dressing*’). Other physician, however, described this step as ‘*Place kerlex under the wrist to place it in a hyper extended position*’. Both these instructions, however, were different from the image’s ground truth (‘*Support patient’s wrist with rolled-up gauze or towel for improved wrist extension*’). Nevertheless, it could be argued that they convey a similar instruction at different levels of detail. This exemplifies the challenge of obtaining a consistent instruction based on medical images. In fact, the average BLEU 1-gram score between the physicians’ instructions was of 0.41, which was lower than the scores obtained when comparing the AI-Medic’s predicted instructions against the ground truth instructions. A possible approach to alleviate these discrepancies is to expand the number of ground truth instructions that describe each image. This would expand the number of correct alternatives and words used to describe each step in a procedure. Likewise, the approach to evaluate the predicted instructions could be modified from a word-by-word metric (BLEU) to a metric that analyzes the semantic content of the instructions. For instance, Natural Language Processing approaches such as Word Mover’s Distance (Kusner et al. 2015), Smooth Inverse Frequency (Arora et al. 2016), or pre-trained encoders like Google’s Sentence Encoder (Cer et al. 2018) could be used to compare the instructions. In doing so, the model would learn to compare the predicted instructions taking their underlying meaning into consideration.

The relatively low BLEU scores are a limitation of our approach that can be attributed to the complexity of the task. For instance, variations in the medical images provided to the neural network will have an effect in the inferences made by the model. Elaborating, the ConvNet would make wrong predictions if the attending surgeon wore blue gloves instead of green or white gloves, as such an image was never seen by the ConvNet during its training process. Other factors include the position of camera, lighting conditions, and the instruments used to complete the procedure (e.g. scalpel instead of scissors). All these factors can play a major role in the inferences made by the network, as they would lead to significantly distinct images even within same steps of a procedure. Wrong predictions at this stage could mean that the local mentee could receive insufficient or incorrect mentoring, which subsequently could impact the patient’s health and safety. A possible countermeasure for this is expanding the dataset by adding more repetitions per procedure. Such images should include variations in orientation, illumination, type of instrument used, etc. While our *Intra-procedure* testing approach alleviates this limitation, more repetitions can improve the prediction results significantly. Additionally, techniques to generate synthetic data (i.e. generated by a machine learning model) can be used to increase the size of the dataset. For example, a Generative Adversarial Neural network (Frid-Adar et al. 2018) could be trained to create new images and descriptions based on the original data. Overall, these data augmentations techniques could lead to improving the model’s accuracy, which could potentially lead to better mentoring and patient outcomes.

The results of these experiments provide subjective and objective evaluation data for the proposed AI mentoring system. Such algorithms and AI based techniques will enable to study new results, such as assessment of the duration of sustained clinical knowledge, the ability to correctly assess the patient condition, the respective treatment through subjective inter-rate reliability indices, and objective performance metrics (e.g. number of errors). By introducing cognitive systems, AI, spatially-augmented reality and physical interaction into the Operating Room (OR), the proposed research holds the promise to reduce morbidity risks due to lack of subspeciality surgical expertise, and to correct clinical judgement and readiness level. Moreover, we anticipate that our framework can be integrated into telementoring systems that support medical training and skills, especially in rural areas, which face significant challenges in securing subspecialist care. The proposed research brings a drastic change in telementoring through the ability of creating a surrogate AI mentor to assist mentees in a rural area, a forward operating base, or community clinics. We foresee these datasets and algorithms playing a key role in enhancing medical training which, in turn, will improve both initial and sustained medical performance. As such, we expect these datasets and benchmark be applicable to the whole medical continuum, producing a significant improvement on all stages of medical care. There are, however, several ethical aspects that need to be address before the implementation of such an AI platform into a fully commercial product. For instance, FDA approval needs to be acquired before its integration into surgical curricula of the United States. For this, the legal implications of such a platform need to be defined and addressed. This includes, but is not limited, defining who should be responsible in case of a medical malpraxis when using the platform in live patients.

Additional future work includes comparing of our approach against existing methods for AI medical instruction. This is a challenging task since there are currently no available benchmark datasets. The DAISI dataset, in fact, addresses this shortage of datasets for autonomous medical mentoring by including step-by-step demonstrations of surgical procedures from various medical specialities. Nonetheless, a possible approach is to compare our approach against datasets comprised of only radiology images (Jing et al. 2017; Pelka et al. 2019; Pavlopoulos et al. 2019). To do this, a subset of the DAISI dataset comprised of only radiology images should be constructed. However, these datasets are used to train algorithms that describe the content of the images. Instead, our approach generates instructions by learning visual-semantic correspondences between images and instructions. Therefore, a comparison against such datasets would require the input of expert radiologists to either convert the instructions of our dataset into descriptions, or convert the descriptions of other datasets into instructions. Finally, the approach can be further improved by including a visual attention framework into the model, such as the one in Xu et al. (2015). Visual attention would allow to visualise how the model focuses its attention over different regions of the input images. This visualisation could be conveyed via augmented reality, superimposing the regions of interest as a saliency map over the view of the operating field. Such visualisations could provide mentees with details on what areas of the operating field are relevant for each step in the procedures.

5. Conclusion

This work presented initial steps towards the development of the AI-Medic, an intelligent artificial system for autonomous medical mentoring. The system uses an encoder-decoder neural network to predict surgical instructions given the current view of a surgery. The AI-Medic was trained using DAISI, a dataset to train AI algorithms that can act as surrogate surgical mentors. The dataset includes 17,339 colour images and captions that provide step-by-step demonstrations for performing surgical procedures from 20 medical disciplines. To assess our system, the instructions predicted by the the AI-Medic were evaluated using cumulative BLEU scores and input from expert physicians. According to the BLEU scores, the predicted and ground truth instructions were as high as $86 \pm 1\%$ similar. Moreover, expert physicians considered that randomly selected images and their predicted descriptions were related. The results from this work serve as a baseline for future AI algorithms assisting in autonomous medical mentoring.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the funders.

Notes on contributors

Dr. Edgar Rojas-Muñoz completed his doctoral studies in the School of Industrial Engineering at Purdue University. He was member of the Intelligent Systems and Assistive Technologies (ISAT) Laboratory at Purdue University. He completed his Licenciatura in Computer Engineering from the Instituto Tecnológico de Costa Rica. His research interests are on gesture understanding, semiotics, human-computer interaction and augmented reality.

Dr. Kyle Couperus, MD is a board certified emergency medicine physician in Tacoma, Washington. He is affiliated with the Emergency Medicine Department of the Madigan Army Medical Center. He completed this residency in the Madigan Healthcare System, and his medical school in the School of Medicina and Biomedical Sciences of the University at Buffalo. He is also a certified member of the American Board of Emergency Medicine.

Dr. Juan Wachs is an Associate Professor in the School of Industrial Engineering at Purdue University and Adjunct Associate Professor of Surgery at IU School of Medicine. He is the director of the Intelligent Systems and Assistive Technologies (ISAT) Lab at Purdue, and he is affiliated with the Regenstrief Center for Healthcare Engineering. He completed postdoctoral training at the Naval Postgraduate School's MOVES Institute under a National Research Council Fellowship from the National Academies of Sciences. Dr. Wachs received his B.Ed.Tech in Electrical Education in ORT Academic College, at the Hebrew University of Jerusalem campus. His M.Sc and Ph.D in Industrial Engineering and Management from the Ben-Gurion University of the Negev, Israel.

ORCID

Edgar Rojas-Muñoz http://orcid.org/0000-0001-6909-375X

References

Alsharid M, Sharma H, Drukker L, Chatelain P, Papageorghiou AT, Noble JA 2019. Captioning ultrasound images automatically. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Shenzhen, China: Springer. p. 338–346.

Arora S, Li Y, Liang Y, Ma T, Risteski A. 2016. A latent variable model approach to pmi- based word embeddings. Trans Assoc Comput Linguist. 4:385–399. doi:10.1162/tacl_a_00106.

Auloge P, Cazzato RL, Ramamurthy N, de Marini P, Rousseau C, Garnon J, Charles YP, Steib JP, Gangi A. 2019. Augmented reality and artificial intelligence-based navigation during percutaneous vertebroplasty: a pilot randomised clinical trial. Eur Spine J. 29(7):1580–1589.

Bilgic E, Turkdogan S, Watanabe Y, Madani A, Landry T, Lavigne D, Feldman LS, Vassiliou MC. 2017. Effectiveness of telementoring in surgery compared with on-site mentoring: a systematic review. Surg Innov. 24(4):379–385. doi:10.1177/1553350617708725.

Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF, others. 2019. Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task. JBJS. 101(23):e127. Publisher: LWW. doi:10.2106/JBJS.18.01197.

Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, et al. 2018. Universal sentence encoder. arXiv Preprint arXiv:1803.11175.

Cortés U, Annicchiarico R, Urdiales C 2007. Agents and healthcare: usability and acceptance. In: Agent Technology and e-Health. pp. 1–4. Birkhäuser Basel.

de Araújo Novaes M, Basu A 2020. Disruptive technologies: present and future. In: Fundamentals of Telemedicine and Telehealth. Academic Press. p. 305–330.

Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 321:321–331. doi:10.1016/j.neucom.2018.09.013.

Geng N, Xie X, Zhang Z. 2019. Addressing healthcare operational deficiencies using stochastic and dynamic programming. Int J Prod Res. 57 (14):4371–4390. Publisher: Taylor & Francis. doi:10.1080/00207543.2017.1397789.

Greenberg CC, Ghousseini HN, Quamme SRP, Beasley HL, Wiegmann DA. 2015. Surgical coaching for individual performance improvement. Ann Surg. 261(1):32–34. doi:10.1097/SLA.0000000000000776.

Hossain MZ, Sohel F, Shiratuddin MF, Laga H. 2019. A comprehensive survey of deep learning for image captioning. ACM Comput Surv (CSUR). 51 (6):1–36. Publisher: ACM New York, NY, USA.

Jha S, MB N. 2019. The Essence of the surgical navigation system using artificial intelligence and augmented reality. Int Res J Eng Technol. 6(4).

Jing B, Xie P, Xing E. 2017. On the automatic generation of medical imaging reports. arXiv Preprint arXiv. 1711.08195.

Karpathy A, Fei-Fei L 2015. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition.. Boston, MA. p. 3128–3137.

Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv Preprint arXiv. 1412.6980.

Kisilev P, Sason E, Barkan E, Hashoul S. 2016. Medical image captioning: learning to describe medical image findings using multi-task-loss CNN. Riva del Garda (Italy): Deep Learning for Precision Medicine.

Kotwal RS, Howard JT, Orman JA, Tarpey BW, Bailey JA, Champion HR, Mabry RL, Holcomb JB, Gross KR. 2016. The effect of a golden hour policy on the morbidity and mortality of combat casualties. JAMA Surg. 151 (1):15–24. Publisher: American Medical Association. doi:10.1001/jamasurg.2015.3104.

Kusner M, Sun Y, Kolkin N, Weinberger K 2015. From word embeddings to document dis- tances. In: International conference on machine learning. Lille, France. p. 957–966.

LeNail A. 2019. Nn-svg: publication-ready neural network architecture schematics. J Open Sour Soft. 4(33):747. doi:10.21105/joss.00747.

Li P, Jiang X, Shatkay H. 2019. Figure and caption extraction from biomedical documents. Bioinformatics. 35(21):4381–4388. doi:10.1093/bioinformatics/btz228.

Lyndon D, Kumar A, Kim J 2017. Neural captioning for the imageCLEF 2017 medical image challenges. CLEF (Working Notes).

Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. 2020. The virtual operative assistant: an explainable artificial intelligence tool for simulation- based training in surgery and medicine. PloS One. 15(2):e0229596. Publisher: Public Library of Science San Francisco, CA USA. doi:10.1371/journal.pone.0229596.

Mishra S, Banerjee M 2020. Automatic caption generation of retinal diseases with self- trained RNN merge model. In: Advanced Computing and Systems for Security. Singapore: Springer. p. 1–10.

Papineni K, Roukos S, Ward T, Zhu WJ 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. Philadelphia, Pennsylvania. p. 311–318.

Patel V, Khan MN, Shrivastava A, Sadiq K, Ali SA, Moore SR, Brown DE, Syed S. 2020. Artificial intelligence applied to gastrointestinal diagnostics: a review. J Pediatr Gastroenterol Nutr. 70(1):4–11. doi:10.1097/MPG.0000000000002507.

Pavlopoulos J, Kougia V, Androutsopoulos I 2019. A survey on biomedical image captioning. In: Proceedings of the Second Workshop on Shortcomings in Vision and Language. Minneapolis, Minnesota. p. 26–36.

Pelka O, Friedrich CM, García Seco de Herrera A, Müller H. 2019. Overview of the image- CLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes; (CEUR Workshop Proceedings. vol. 2380; Lugano, Switzerland: CEUR-WS.

Rojas-Muñoz E, Lin C, Sanchez-Tamayo N, Cabrera ME, Andersen D, Popescu V, Barragan JA, Zarzaur B, Murphy P, Anderson K, et al. 2020. Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. NPJ Digi Med. 3(1):1–9. doi:10.1038/s41746-020-0284-9

Schuster M, Paliwal KK. 1997. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 45(11):2673–2681. doi:10.1109/78.650093.

Sebajang H, Trudeau P, Dougall A, Hegge S, McKinley C, Anvari M. 2005. Telementoring: an important enabling tool for the community surgeon. Surg Innov. 12(4):327–331. Publisher: Westminster Publications, Inc. 708 Glen Cove Avenue, Glen Head, NY 11545, USA. doi:10.1177/155335060501200407.

Sebajang H, Trudeau P, Dougall A, Hegge S, McKinley C, Anvari M. 2006. The role of telementoring and telerobotic assistance in the provision of laparoscopic colorectal surgery in rural areas. Surg Endosc Other Interven Techn. 20(9):1389–1393. Publisher: Springer. doi:10.1007/s00464-005-0260-0.

Simonyan K, Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv. 1409.1556.

Singh S, Karimi S, Ho-Shon K, Hamey L 2019. From chest X-rays to radiology reports: a multimodal machine learning approach. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). Perth, Australia: IEEE. p. 1–8.

thumbroll LLC. 2020. Thumbroll. [accessed 2020 Mar 13]. https://www.thumbroll.com/ .

Ward KPSRT, Reeder JHF 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results.

Xu J, Liu W, Liu C, Wang Y, Chi Y, Xie X, Hua X 2019. Concept detection based on multi- label classification and image captioning approach-damo at imageclef 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings,(CEUR-WS. org). Lugano, Switzerland. p. 1613–1673.

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y 2015. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning. Lille, France. p. 2048–2057.