Producing radiologist-quality reports for interpretable artificial intelligence.

William Gale, Gustavo Carneiro

School of Computer Science
The University of Adelaide
Adelaide, SA 5000
will@wgale.com
gustavo.carneiro@adelaide.edu.au

Luke Oakden-Rayner, Lyle J. Palmer

School of Public Health
The University of Adelaide
Adelaide, SA 5000
{luke.oakden-rayner,lyle.palmer}
@adelaide.edu.au

Andrew P. Bradlev

Faculty of Science and Engineering Queensland University of Technology Brisbane, QLD 4001 a6.bradley@qut.edu.au

Abstract

Current approaches to explaining the decisions of deep learning systems for medical tasks have focused on visualising the elements that have contributed to each decision. We argue that such approaches are not enough to "open the black box" of medical decision making systems because they are missing a key component that has been used as a standard communication tool between doctors for centuries: language. We propose a model-agnostic interpretability method that involves training a simple recurrent neural network model to produce descriptive sentences to clarify the decision of deep learning classifiers. We test our method on the task of detecting hip fractures from frontal pelvic x-rays. This process requires minimal additional labelling despite producing text containing elements that the original deep learning classification model was not specifically trained to detect. The experimental results show that: 1) the sentences produced by our method consistently contain the desired information, 2) the generated sentences are preferred by doctors compared to current tools that create saliency maps, and 3) the combination of visualisations and generated text is better than either alone.

1 Introduction

Recent advances in machine learning techniques have resulted in medical decision making systems that equal human doctors at a variety of tasks [1, 2], but significant barriers to clinical implementation remain. One widely discussed issue in the machine learning community, has been the concept of interpretability [3] – how believable and understandable the decisions of machine learning models are. Indeed, the European Union has recently passed legislation requiring machine learning systems to be explainable if their decisions affect humans [4]. In response to these concerns, researchers have sought to "open the black box of AI" [5] by attempting to explain the decisions of systems in a variety of ways, which typically have fallen into the following categories: 1) identifying regions of the image that contribute to the decision (saliency mapping/heatmaps) [6], 2) visualising the learned features of deep neural networks [7], and 3) identifying clusters of samples that receive the same decisions, for human exploration [8].

^{*} These authors contributed equally to the work

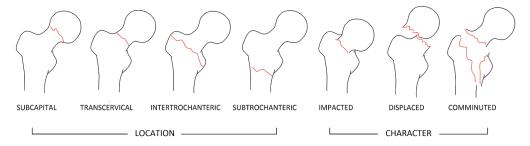


Figure 1: Key terms describing the location and character of hip fractures.

While these methods provide some insights into the inner workings of deep neural networks, they have yet to convincingly address concerns around model interpretability. We believe that this is because there are two competing broad definitions of interpretability. Machine learning specialists seek to understand the mechanisms underlying their models, but doctors and other end-users simply want "human-style" explanations; the same sort of explanations they currently receive from other humans. Humans explain their decisions with natural language descriptions, either in the form of speech or a written report. While this description is often post-hoc and may not accurately reflect the decision making process, it is the form of explanation that end-users currently expect and are satisfied with. For example, if a human radiologist says "there is an irregular mass with abnormal blood flow that is consistent with cancer" this description both communicates the relevant finding, but also describes the important image features that informed that decision. The clinician receiving the report can judge whether the description matches the decision, and whether the description matches their own interpretation of the image.

There have been multiple previous attempts to produce descriptive text from natural images [9, 10] and radiology reports [11, 12], but these methods were employed to make use of readily available free-text data as a training signal with the goal of improving model performance. In the case of the radiology tasks, attempting to reproduce highly variable radiology reports resulted in outputs that appear too inconsistent to be useful as explanations. Even in the most cutting edge work in the field [12] the text produced is far from human-like and would not satisfy doctors as either a diagnostic report, or as an explanation for model decisions.

In this paper, we are motivated by the idea that the explanation given by a doctor for any narrow medical task is often fairly simple, and the variability in the phrasing of these reports does not change their meaning. Instead of trying to reproduce whole reports, or even just the relevant report sentences, we attempt to train an image-to-text model to produce simplified but meaningful descriptions that describe and justify a diagnostic decision in a way that is satisfying to clinicians.

In this work we present a model-agnostic extension of deep learning classifiers based on a recurrent neural network model with a visual attention mechanism. This model produces a short text description to explain the medical decisions made by the classifier. We test our tool on a large dataset of hip fractures in pelvic x-rays, and show that our model can produce high quality "human-style" explanations.

2 Materials and Methods

2.1 Materials

The dataset available for this work consists of 50,363 frontal pelvic X-rays, containing 4,010 hip fractures [13]. These images were randomly divided into a training set (41,032 images, with 2923 hip fractures), a validation set (4,754 images, with 414 hip fractures) for model selection, and a held-out test set (4,577 images, with 348 hip fractures). There was no overlap of patients between the three sets. The deep learning classifier explored in our paper is the recently developed DenseNet [14], trained to classify hip fractures from frontal pelvic X-rays [13]. This model has been shown to produce a classification performance equivalent to a human radiologist with an area under the ROC curve of 0.994 [13].

The original training dataset for the CNN consisted of images with labels which included the anatomical location of the abnormality (Figure 1). We also had access to descriptive sentences for each fracture retrieved from the original radiology reports, but these had highly inconsistent struc-

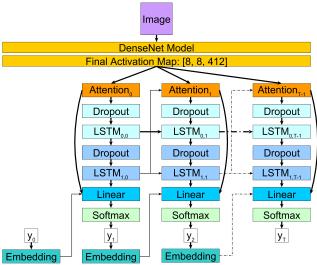


Figure 2: The Architecture of the Image-to-Text Model

ture and content, making the task of training a system to generate similar sentences very difficult. To simplify the task, a radiologist created a new set of hand-labelled descriptive terms based on the image appearances. The elements of a radiology report we considered important in the diagnosis of hip fractures were the degree of fracture displacement, the degree of fragmentation (comminution), and the presence of subtle impaction (where one part of the bone is pushed into another), also shown in Figure 1. Unlike the task of identifying the presence or absence of fractures (a difficult task for humans [15]), describing the visual features of a known fracture is trivial. For instance, the radiologist was able to label all of the 4,010 fractures with these descriptive terms (a total of 7 labels) in under 3 hours.

Because of the variation in language and quality in the original sentences, we use categorical variables for location and character to generate new sentences with a standard grammar and limited vocabulary. Each sentence had the general structure: "There is a [degree of displacement], [+/-comminuted][+/- impacted] fracture of the [location] neck of femur [+/- with an avulsed fragment]." Negative cases (i.e., those without fractures) were automatically labelled with a consistent sentence - "No fracture was identified on this study". This sentence set was used to train an interpretability tool (described below), with the weights of the DenseNet held fixed. Therefore the image-analysis part of the model was never re-trained on the descriptive terms in the sentences (for example, to distinguish fractures with mild or severe displacement). We hypothesise that features relevant to this task were learned in the process of producing a fracture-detecting model.

2.2 Model

Let us define the dataset as $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i,\mathbf{a}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i:\Omega\to\mathbb{R}$ denotes an X-ray image (with Ω representing the image lattice), $\mathbf{y}_i=[\mathbf{y}_i(1),\ldots,\mathbf{y}_i(T)]$, with $\mathbf{y}_i(t)\in[0,1]^K$ - that is, \mathbf{y}_i is a report containing T words and the vocabulary size is K (in this paper K=30, where the vocabulary comprises 26 lower-cased words plus start (SOT), end (EOT), unknown (UNK) and pad tokens), and $\mathbf{a}_i=[\mathbf{a}_i(1),\ldots,\mathbf{a}_i(C)]$, with $\mathbf{a}_i(c)\in\mathbb{R}^D$ ($C=8\times 8$ represents the number of image regions and D=412 denotes the region representation size), is the representation obtained from the trained DenseNet [13], using the final activation map, i.e., the layer preceding average pooling and softmax.

Our proposed model agnostic interpretability method is defined by a recurrent neural network (RNN) model composed of two long short term memory (LSTM) [16] layers, as depicted in Fig.2, where the LSTM layers are defined by

$$\mathbf{h}_{0,t} = LSTM_0(\mathbf{z}_{0,t}, \mathbf{h}_{0,t-1}), \mathbf{h}_{1,t} = LSTM_1(\mathbf{z}_{1,t}, \mathbf{h}_{1,t-1}),$$
(1)

where $\alpha_t = [\alpha_{t,1}, \dots, \alpha_{t,C}]$ (with $\alpha_{t,c} \in [0,1]$) denotes a soft attention vector to be applied to each one of the C regions of \mathbf{a} and is estimated from inputs $\mathbf{h}_{1,t-1}$ and \mathbf{a} , $\mathbf{z}_{0,t} = \text{dropout}(\alpha_t \odot \mathbf{a})$ (we are abusing the notation for the elementwise multiplication \odot by allowing each $\alpha_{t,c}$ to multiply the whole D-dimensional vector $\mathbf{a}(c)$, and dropout is defined in [17]), $\mathbf{z}_{1,t} = \text{dropout}(\mathbf{h}_{0,t})$ and $\mathbf{h}_{0,t}, \mathbf{h}_{1,t} \in \mathbb{R}^{512}$ are the hidden state representations. During inference, the initial hidden states are defined as $\mathbf{h}_{0,0} = \mathbf{h}_{1,0} = \vec{0}$, the initial word $\mathbf{y}(0)$ is a special SOT token and $\mathbf{y}(T)$ is the special EOT token. The output layer [18], which predicts the next word in the sequence is defined by:

$$\widetilde{\mathbf{y}}_t = \operatorname{softmax}(f_{\mathbf{W}}(\mathbf{h}_{1,t}, f_{\mathbf{E}}(\mathbf{y}_{t-1}), \boldsymbol{\alpha}_t \odot \mathbf{a})),$$
 (2)

where **W** and **E** are learnable parameters of the linear functions $f_{\mathbf{W}}(.)$ and $f_{\mathbf{E}}(.)$, respectively, with $f_{\mathbf{E}}(.)$ used to calculate a 256-dimensional word embedding. Both LSTM layers in (1) have learnable parameters [16], where the training process uses the length normalised cross-entropy loss, defined as follows for each training sample:

$$\ell = \frac{\sum_{t=1}^{L} \mathbf{y}_{t}^{T} \log(\widetilde{\mathbf{y}}_{t})}{L}.$$
(3)

The training is regularised using dropout with an 80% keep-rate and data augmentation of the input image during training.

3 Experiments and Results

3.1 BLEU Score

The BLEU score [19] measures the similarity between the reference and model-generated sentences on the test set for 1-gram, 2-gram, 3-gram, and 4-gram. We calculated the BLEU score for the entire set of fractures in the held-out test set (348 images). To demonstrate the effect of simplifying our explanatory sentences, we show results for the same model architecture trained on either the original reports or our new sentences. We did not include the non-fracture test cases since the automatically generated "null" sentence would be flawlessly reproduced and identical for these samples, inflating the score. The results (Table 1), showed a high level of reproduction accuracy. While these results are unsurprising (i.e., that an RNN can easily produce sentences with a simple structure and small vocabulary), they do confirm the hypothesis that these simplified sentences are highly learnable, despite the fact that CNN component was not trained on the descriptive terms in these sentences.

Table 1: BLEU scores on the fractures in the test set, comparing models trained to reproduce the original report sentences vs our simplified explanatory sentences.

1 <u> </u>		
	Original reports	Simplified sentences
1-gram	65.0	91.9
2-gram	37.9	83.8
3-gram	24.2	76.1
4-gram	15.9	67.7
Weighted average	25.67	77.97

3.2 Sentence Content

The radiologist who created the original labels reviewed the sentences and images for 200 randomly chosen fractures from the test set in order to assess the semantic content of the sentences produced by our model. The radiologist reported the percentage of model-produced sentences and sentences from the original radiology reports that contained the appropriate location and character terms to describe the fractures.

The text generation method was slightly worse at identifying the appropriate location, but better at describing the characteristics of the fracture (Table 2).

The most common error in the generated sentences were "off-by-one" location errors, for example describing a subcapital fracture as transcervical. The most common error in the original radiology reports was failing to describe the character of the fracture.

Table 2: The percentage of explanatory sentences that contain an appropriate description of the location and character of the fractures, as determined by a radiologist.

	Appropriate description	Appropriate description
	of location	of fracture character
Original radiology reports	99%	78%
Our generated sentences	90%	98%

3.3 Acceptance of Explanations by Doctors

We presented several forms of explanations to 5 doctors, all with between 3 and 7 years of post-graduate clinical experience. We selected a subset of 30 cases from the test set, comprising 10 randomly selected fractures from the groups with mild, moderate, and severe displacement (to ensure a variety of examples). For each of these cases, we present 3 variations for explanations: 1) saliency maps produced by the SmoothGrad [6] method, 2) sentences generated by our model, and 3) a combination of saliency maps and sentences. Each doctor was asked to assume they had received the diagnosis for each image from an unknown and untrusted source. They were asked to score each type of explanation on a Likert scale of 1 to 10, where 1 reflects a completely unsatisfactory explanation, and 10 reflects a perfect explanation. The average scores for each method are presented in Table 3, revealing that doctors prefer "human-style" text explanations over saliency maps, and have a preference for a combination of both saliency maps and generated text together compared to either alone. While the scale was arbitrary, the high scores consistently given to the combination of text and saliency maps suggest that the doctors were most satisfied with these type of explanations.

Table 3: The average qualitative score and the range of scores given by a group of 5 doctors for different explanations for a diagnosis of hip fracture.

	Average qualitative score (range)
Saliency maps alone	4.4 (2-6)
Generated sentences alone	7.0 (6-8)
Saliency maps and sentences	8.8 (8-9)

Finally, we present randomly selected examples of fractures, saliency maps, and both the original report descriptions and the generated text explanations (Figure 3). Further examples are provided in the supplementary material.

4 Discussion

We have shown experimentally that doctors prefer our generated text over saliency maps produced by a popular visualisation method, and prefer a combination of both over the generated sentences alone.

Our method is model-agnostic, meaning it can be applied to any deep learning image analysis model. We argue that a well-trained CNN that demonstrates human-level or greater performance at a medical task will already have learned to identify features to explain its decisions, and can be extended to produce a constrained text output for this task. Distilling explanatory sentences to their important elements makes the task tractable. The time cost of additional labelling appears to be manageable, at least in comparison to the effort of building the decision-making model. We believe this is because the descriptive terms are usually more easily identified and labelled than standard diagnostic tasks, and the labels required to build the diagnostic model will often overlap with the labels required to generate explanatory text. We expect this to be true in many medical settings; most individual image-based diagnoses in fields such as radiology, dermatology, and pathology are informed by very few key positive and negative findings, allowing for a similar approach to explanatory text generation.

5 Conclusion

Interpretable decision making is necessary in medicine, and machine learning systems that make medical decisions will be expected to address this need. Given that human-produced reports are considered sufficient as a form of explanation for medical decisions, we believe that our approach to generate text-based explanations provides a path to harmonise machine learning decisions and the human demand for explainability.

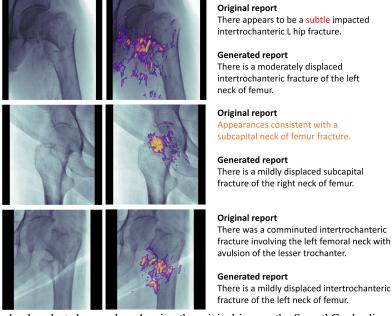


Figure 3: Randomly selected examples, showing the original image, the SmoothGrad saliency map [6], the original report description, and the generated sentence. Orange highlights indicate that the sentence does not appropriately characterise the fracture. Red highlights indicate that the radiologist thought an alternative descriptive term was more appropriate.

References

- [1] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama **316**(22) (2016) 2402–2410
- [2] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639) (2017) 115
- [3] Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
- [4] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a" right to explanation". arXiv preprint arXiv:1606.08813 (2016)
- [5] Castelvecchi, D.: Can we open the black box of ai? Nature News 538(7623) (2016) 20
- [6] Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
- [7] Mordvintsev, A., Olah, C., Tyka, M.: Inceptionism: Going deeper into neural networks. Google Research Blog. Retrieved June **20**(14) (2015) 5
- [8] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov) (2008) 2579–2605
- [9] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5) (2016) 1285–1298
- [10] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. (2015) 2048–2057
- [11] Krening, S., Harrison, B., Feigh, K.M., Isbell, C.L., Riedl, M., Thomaz, A.: Learning from explanations using sentiment and advice in rl. IEEE Transactions on Cognitive and Developmental Systems **9**(1) (2017) 44–55

- [12] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. arXiv preprint arXiv:1801.04334 (2018)
- [13] Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A.P., Palmer, L.J.: Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv preprint arXiv:1711.06504 (2017)
- [14] Huang, G., Liu, Z.: Densely connected convolutional networks. In: CVPR. (2017)
- [15] Cannon, J., Silvestri, S., Munro, M.: Imaging choices in occult hip fracture. Journal of Emergency Medicine **37**(2) (2009) 144–152
- [16] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
- [17] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1) (2014) 1929–1958
- [18] Pascanu, R., Gulcehre, C., Cho, K., Bengio, Y.: How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026 (2013)
- [19] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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 (2002) 311–318