Natural Language Generation Model for Mammography Reports Simulation

Assaf Hoogi, Arjun Mishra, Francisco Gimenez, Jeffrey Dong, and Daniel Rubin

***Abstract*—Extending the size of labeled corpora of medical reports is a major step towards a successful training of machine learning algorithms. Simulating new text reports is a key solution for reports augmentation, which extends the cohort size. However, text generation in the medical domain is challenging because it needs to preserve both *content* and *style* that are typical for real reports, without risking the patients' privacy. In this paper, we present a conditioned LSTM-RNN architecture for simulating realistic mammography reports. We evaluated the performance by analyzing the characteristics of the simulated reports and classifying them into benign and malignant classes. An average classification AUC was calculated over two distinct test sets. A qualitative analysis was also performed in which a masked radiologist classified 0.75 of the simulated reports as real reports, showing that both the style and content of the simulated reports were similar to real reports. Finally, we compared our RNN-LSTM generative model with Markov Random Fields. The RNN-LSTM provided significantly better and more stable performance than MRFs ($p < 0.01$, Wilcoxon).**

***Index Terms*—Natural language generation, mammography reports, RNN-LSTM, simulation.**

I. Introduction

USING machine-learning tools for medical applications has become very popular over the last few years, but progress towards models that are useful in practice has been hindered by a dearth of annotated clinical data. Obtaining annotated data is often challenging, time consuming and expert-dependent as well. Generation of medical text reports is highly valuable because annotated medical data is not commonly available. In addition, if our small cohort of real reports is de-identified, then using it as a training set and generating many other de-identified reports will enrich the data while minimizing the risk for the patient privacy. Several works contribute to large-scale de-identification of datasets using techniques such as generalization, suppression (removal), or permutation and swapping of certain data values [1], [3], [4], [8], [13], [16], [18], [32], [37]. In the medical informatics community, there are many efforts to de-identify medical text documents, dealing with HIPAA identifiers [19], or Protected Health Information (PHI) [26]. Szarvas *et al.* developed a model for anonymizing Protected Health Information (PHI) while Shweta *et al.* presents another method for de-identifying medical records using an Recurrent Neural Network (RNN) [30]. However, Sweeney *et al.* shows that de-identified datasets are nonetheless subject to the risk of re-identification of those patients [29], by tracing back to the imaging machine. Therefore, rather than using de-identification procedure, generation of simulated data could be an optimal alternative strategy for extending medical datasets. This can be used for training machine learning algorithms or radiologists trainees in a manner that provides the important statistical properties of actual data *without risking patient privacy*.

II. Related Work

Many works to date have used machine-learning techniques for text generation over different domains, such as speech-to-text applications and image captioning (image-to-text). Natural Language Generation (NLG) models, either rule-based or corpus-based, can be used for these goals. Rudnicky and Oh, for instance, proposed an n-gram language model approach [23] while Mairesse and Young proposed a phrase-based NLG system based on factored language models that can learn from a semantically aligned corpus [20]. Karpathy *et al.*, on the other hand, presented a technique for captioning different regions in images [15]. Their approach leverages datasets of images combined with their text descriptions to learn about the inter-modal correspondences between natural language and visual data. Their model is based on a combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective approach that aligns the two modalities through a multimodal embedding. Socher *et al.* [27] further studied the image-text correlation to annotate different parts of images. Several works generate image captions by using fixed templates that are filled based on the content of the image [11], [17], [33], [34], but this approach limits the variety of possible outputs. In recent work, Dong *et al.* propose a new training method called Image-Text-Image that integrates text-to-image and image-to-text synthesis to improve the performance of text-to-image synthesis [7]. However, despite the substantial focus on text generation across multiple domains, there are only a few works that focus on

Manuscript received July 26, 2019; revised November 20, 2019 and January 28, 2020; accepted March 1, 2020. Date of publication April 20, 2020; date of current version September 3, 2020. This work was supported in part by a grant from GE Medical Systems under the GE Healthcare Blue Sky program. *(Corresponding author: Assaf Hoogi.)*

Assaf Hoogi, Francisco Gimenez, Jeffrey Dong, and Daniel Rubin are with Biomedical Data Science, Stanford University, Stanford, CA 94305 USA (e-mail: assafho@weizmann.ac.il; francisco.gimenez@gmail.com; jefdongus@gmail.com; dlrubin@stanford.edu).

Arjun Mishra is with Computer Science, Berkeley City College, Berkeley, CA 94704 USA (e-mail: arjunsaxmishra@gmail.com).

Digital Object Identifier 10.1109/JBHI.2020.2980118

2168-2194 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.

generative models for *text-to-text* simulation. By extending an Long Short-Term Memory (LSTM) network to be both deep in space and time, Graves shows that the resulting network can be used to synthesize handwriting indistinguishable from that of a human [10]. Zhang and Lapata use RNNs to generate Chinese poetry [35]. In [36], Zhang *et al.* present an automatic drawing of Chinese characters by using an LSTM conditional generative model. Their technique was used to model Chinese handwriting, allowing the method to generate new handwritten characters by sampling from the probability distribution associated with the RNN [36]. A two-component method was proposed by Cho *et al.* [5] and Sutskever *et al.* [28] to encode a variable-length source sentence into a fixed-length vector and to decode the vector into a variable-length target sentence. Encoding to a fixed-length vector has its limitations; thus, Bahdanau *et al.* proposed a modified method without this constraint [2]. Despite the advances of prior works within the text generation domain, a key challenge for text-text generation is the simulation of new text that not only preserves the style of the original text, but also the meaningful content of the text. For example, Karpathy demonstrated a character-level RNN architecture for generating Shakespeare text. The RNN architecture did well in replicating the typical Shakespeare style, but the system had substantial difficulties in generating content with reasonable meaning. On contrary, the challenge of generating new text reports, preserving the style-content characteristics of real ones, is extremely relevant in the medical domain.

This paper presents several key contributions.

- We propose a Natural Language Generation model for text-to-text simulations. This model preserves the content-style characteristics of the original text. The reports' simulation is important for data enrichment task, which extends the training set by adding PHI protected reports. To our knowledge, a successful content-style NLG model has not previously been reported.
- This is the first method that was developed for generation and augmentation of medical text reports, and specifically for radiology reports. The medical domain has unique NLG challenges: the generated report must satisfy the constraints of *grammaticality* (relevant subsections, language style and punctuations), *meaningfulness* (sentences meaning, clinical diagnosis) and *reasonableness* in a medical sense (i.e., a medical report must exhibit features that distinguish it from non-medical text). In addition, there is an inherent diversity of medical reports, as each clinician has their own particular way of describing the same diagnosis, spelling or grammar errors are common, and there inevitably exist misdiagnoses or incomplete interpretations in any clinical report corpus each of these should be handled by a robust text generation technique.
- We present a comparison of the LSTM model with an alternative approach for generating simulated reports with a Markov Random Field (MRF) approach, and we quantitatively and qualitatively evaluate performance of each technique.

The paper is organized as follows. Section II introduces the NLG model for end-to-end simulation of mammogram reports. Section III describes the experimental dataset, Section IV presents the results, and Section V draws the concluding remarks about the quality of the presented method.

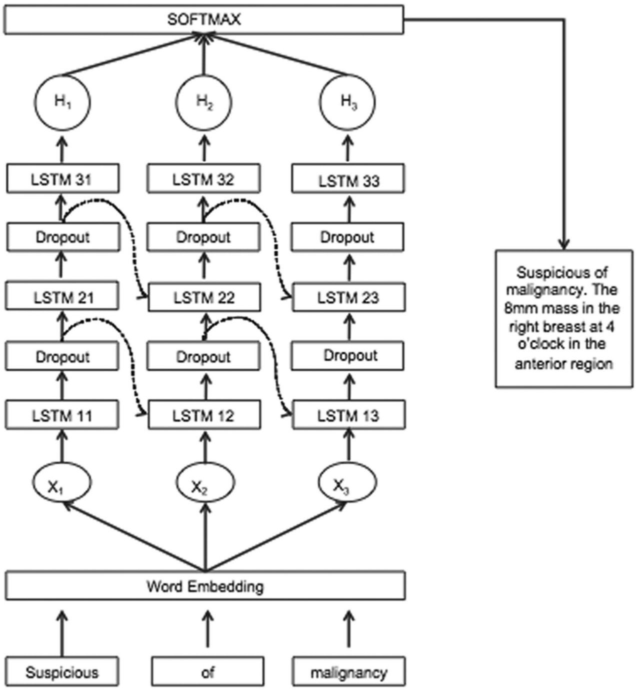


Fig. 1. Our RNN-LSTM architecture for text-to-text report generation.

III. The Proposed Method

In this work, we use a 3-layer RNN-LSTM architecture as a NLG tool for text-to-text simulation (Fig. 1). RNN-LSTM-based generative models are commonly used for both unconditional [12], [14] and conditional text [31] generation. These models represent a class of autoregressive models that generate outputs sequentially, where the next predicted element is conditioned or unconditioned on the previous generated elements [25]. For our model, we chose the conditional text generation approach.

A. Conditional Generative Text-to-Text RNN-LSTM Model

Text-to-text generation aims to directly model the conditional probability $p(y|x)$ of mapping an input sequence $(x_1,..,x_T)$ into an output sequence $y_1,..,y_{T'}$ whose length $T^{'}$ may be different than T. We compute the conditional probability by first obtaining the fixed dimensional representation of the input sequence, given by the last hidden layer of the LSTM, and then compute the probability of the output with a standard LSTM formulation by the following

$$p(y_1,..,y_{T'},x_1,..,x_T) = \prod_{t=1}^{T^{'}} -(y_t|\eta, y_1.y_{(t-1)}) \qquad (1)$$

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.

Let the memory cells $(h_1, .., h_T)$ in the LSTM layer produce a representation sequence. This representation sequence is then averaged over all time steps resulting in representation h. Let (C_t) be a candidate for the states of a memory cell at time t,

$$\tilde{C}_t = tanh(W_c x_t + U_c h_{(t-1)} + b_c) \tag{2}$$

where W_c is the input-to-hidden weight matrix, U_c is the state-to-state recurrent weight matrix, and b_c is the bias vector. Then, the memory cells new state can be computed by:

$$C_t = i_t \otimes \tilde{C}_t + f_t \otimes C_{(t-1)} \tag{3}$$

where the operation $\otimes$ denotes the element-wise vector product. Output gates $(o_1, .., o_T)$ can then be calculated as follows:

$$o_t = \sigma(W_0 x_t + U_0 h_{(t-1)} + V_0 C_t + b_0) \tag{4}$$

$tanh$ is used to update the sequence h_T that is represented as:

$$h_t = o_t \otimes tanh(C_t) \tag{5}$$

The forget gates of the LSTM that we use in this work are controlled by the sigmoid function:

$$f_t = \sigma(W_f x_t + U_f h_{(t-1)} + b_f) \tag{6}$$

We use a dropout step after each LSTM layer as well as the RMSPROP optimizer [6]; we also utilized the Softmax function over all output vectors together.

B. Word Embedding

Word embedding methods have the powerful capability to capture both semantic and syntactic variations of words [21]. In this work, we use GloVe (Global Vectors for Word Representation) model [24] that is an unsupervised learning algorithm for obtaining vector representations. We train the GloVe on Wikipedia text and the iterative learning is done through sampling the word co-occurrence distribution. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. We use a skip-gram with a negative sampling in our model [9], [22]. As long as the underlying dataset is being bootstrapped from is PHI protected real reports, there are no PHI concerns in the simulated data. Therefore while the actual GloVe embeddings from the Wikipedia corpus may contain arbitrary names, as a result of the underlying training data they will never be encoded.

C. Model Training

Our LSTM architecture was trained by using the whole MCW cohort (see Experimental Dataset section). In order to best learn the statistics of each type of reports (benign/malignant), we designed two parallel word-level LSTM architectures, for benign and malignant cases *separately*, with three layers each. The main reason for separating the reports is the size of each dataset. The malignant cohort is significantly smaller than the benign one, and as a result the architecture hyper-parameters must be different in order to obtain optimal performance. By learning the statistics of the benign and malignant reports separately, we can utilize the entire cohort without explicitly addressing class imbalance within the model construction. Hyper-parameters of the architecture were optimized by applying a grid search, wherein different combinations of parameters have been tested and the best one was chosen. As a result of the difference in the benign and the malignant cohort sizes, we selected 600 neurons and batch size of 100 in each LSTM layer for the benign subset and 200 neurons and batch size of 20 in each LSTM layer for the malignant reports architecture. We set the learning rate to 0.002 and use an adaptive learning rate method, RMSProp, with a decay rate of 0.97. A dropout of 0.2 to reduce over-fitting was applied.

D. Reports Evaluation

To evaluate the quality of the simulated reports, we used both qualitative and quantitative analyses.

E. Qualitative Analysis

For the qualitative analysis, 30 different reports were randomly selected. Fourteen reports were real, 16 were simulated. Fiftheen were malignant and others (15) were benign. A radiologist who was unaware of the type of each report (simulated or real) was asked to classify the reports into these two classes – this procedure was performed in order to assess how similar the simulated reports were to the real ones with respect to both content and style. Each report has two separate sections: the $findings$ section that records the imaging observations and the $impression$ section that records the overall conclusion/diagnosis. The radiologist made his decision by exploring the findings consistency and their fit to the final report's conclusion. The radiologist checked the grammar (relevant subsections, language style and punctuations), the meaning (sentences meaning, clinical diagnosis) and its reasonableness in a medical sense.

F. Quantitative Analysis

1) ***Report Classification Using Bayesian Network****:* We first extracted the imaging findings from the free-text simulated reports. The extracted observations were input into a Bayesian Network (BN) that was used to classify the simulated reports into benign and malignant cases by predicting the likely diagnosis based on the input imaging findings. We then compared the concluded diagnosis reported in the impression section of each simulated report against the most likely diagnoses that we derived from the BN model. This was done to evaluate the internal consistency of the simulated reports and the correlation between the $findings$ and the $impression$ sections in each report. In this work, the BN model was used for reports' evaluation *only*. A key reason for choosing BN is that its degrees of freedom are relatively small. Therefore, a variation between two Bayes Nets is presumably due to the data rather than any variation in the Bayes Nets. This is because the Bayes Net is just learning Conditional Probability Tables which are inherently related to the data, it does not learn other parameters. The training of the BN model was done by using the MCW dataset, from which we generated a subset of 984 cases (492 from each class). We

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.

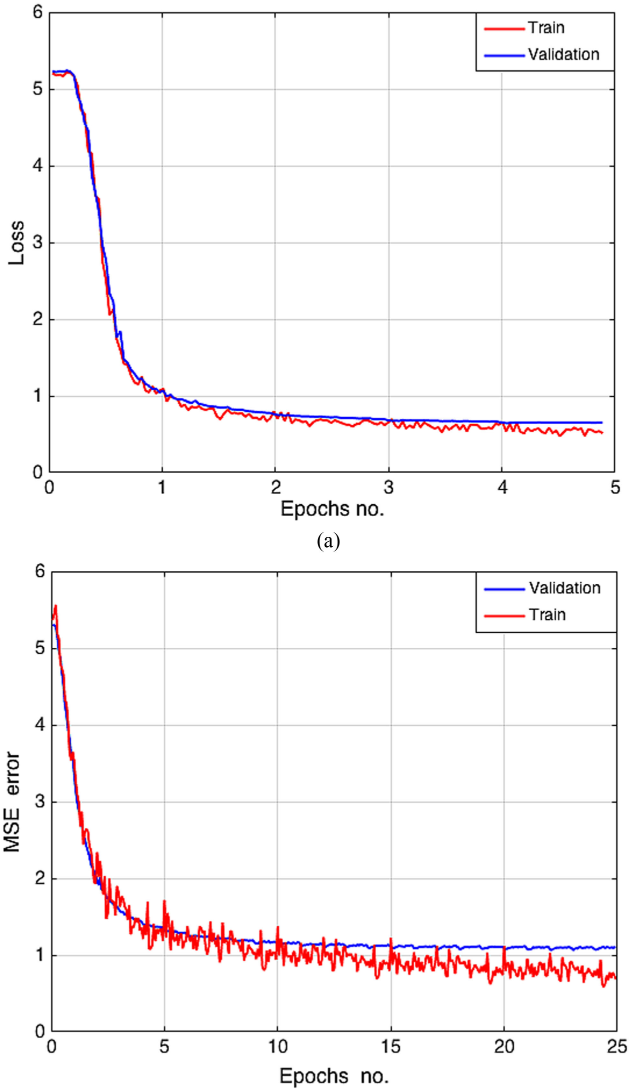


Fig. 2. Training and validation Mean Squared Error for both (a) benign and (b) malignant cases.

then used 20-fold cross-validation and measured the area under the curve (AUC) and Gini coefficient. Gini coefficient applies to binary classification and requires a classifier that can in some way rank examples according to the likelihood of being in a positive class. Both AUC and Gini coefficient were computed using the BN probability of a lesion being malignant or benign compared to the actual lesion type that was obtained by the impression section in the reports.

*2) **Exploring Over-Fitting**:* Because we did not have many malignant cases, over-fitting is a possibility. In order to reduce the chance of over-fitting, our architecture 1) includes both batch normalization and dropout regularization, 2) uses training and testing subsets that were chosen from different data sources (MCW/Stanford), to prevent any over-fitting to a specific dataset, 3) is relatively shallow, which decreases the risk of over-fitting. Encouragingly, we observed that both training and testing errors were comparable (Fig. 2). For the malignant cases, we can see that 1) there is a slight difference between the training and the validation loss, 2) the validation loss is a bit noisy. This finding is reasonable considering the small number of malignant reports that we had. However, this difference is not significant (Wilcoxon text, $p > 0.05$), and this is not considered to be an instance of over-fitting.

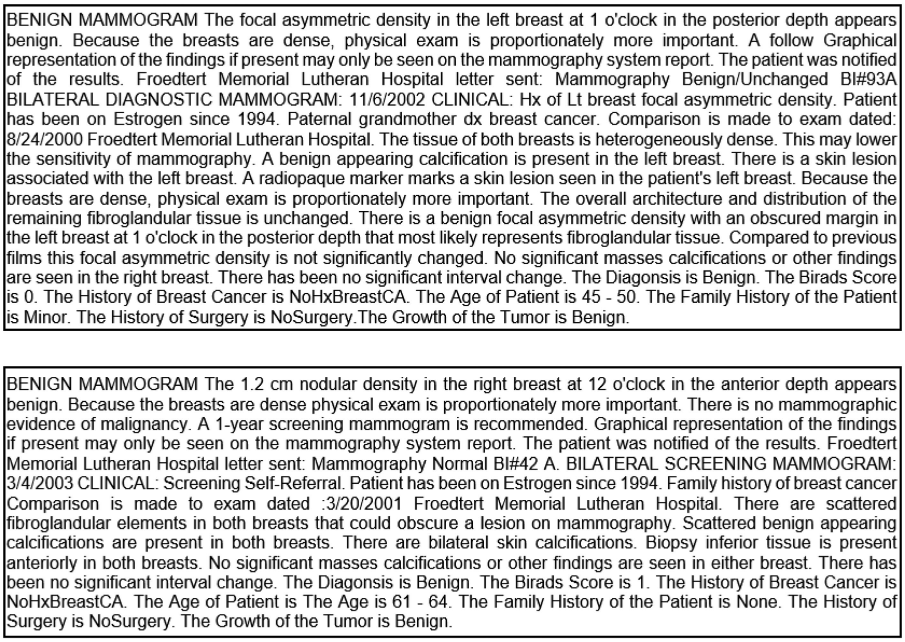
*3) **Contribution of Simulated Reports as a Data Augmentation Tool**:* As was mentioned, the Bayesian Network is not an integrative part of the NLG model. In this work, it was used *only* as an evaluation tool for two different purposes: 1) to evaluate internal consistency and accuracy of the report generation, and 2) to evaluate the contribution of the simulated reports as a data augmentation approach. We conducted two different experiments. *First*, we used a *varied* total number of training examples. We explored the value of adding a varied number of simulated reports (0, 30, 50, 100) to a fixed number of 50 real reports. The BN was then used to classify a separate test set of real reports into benign and malignant classes. *Second*, we used a *fixed* total number of training examples. Then, we changed the percentage of the real and the simulated reports within the whole cohort - 1) real reports only, 2) 50% real reports and 50% simulated ones, and 3) 25% real reports and 75% simulated ones.

*4) **Comparison With Markov Random Field Generative Model**:* We compared our RNN-LSTM generative model with a Markov Random Fields (MRF) generative one, a computationally efficient technique that can be used for text generation. MRFs do not require significant computational resources, however, they are unable to capture long-range dependencies in the same manner as RNN-based architectures [34]. To compare these two approaches, we explored the classification accuracy of the generated reports by training the BN classifier with 1) the original data only, and 2) a combined dataset of both real and simulated reports. Gini coefficient and AUC values were calculated between the automated classification and the ground truth (impression section in the reports) for each method in order to evaluate the classification accuracy.

IV. Experimental Dataset

We used a large cohort of mammography narrative reports that includes two different subsets 1) the Medical College of Wisconsin (MCW) subset and 2) the Stanford mammography subset. The MCW dataset was used to train the LSTM architecture during the report generation phase as well as to train and test the Bayesian Network used for evaluation of the test reports. The Stanford dataset was used only as a test set for the Bayesian Network. In that way, we were able to explore overfitting to a specific dataset. The MCW subset includes reports from patients collected at the Froedtert Memorial Hospital and Medical College in Milwaukee. The reports are based on consecutive screened and diagnostic mammography reports in 18,269 patients. There are a total of 61,684 reports in our MCW subset, 492 of which are classified as malignancies. This subset includes reports from all 6 BIRADS diagnostic categories. The Stanford subset is based on patients observed at Stanford Hospital via consecutive screened and diagnostic mammography reports. This dataset contains a total of 107 malignant cases and 37,570 benign cases.

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.



BENIGN MAMMOGRAM The focal asymmetric density in the left breast at 1 o'clock in the posterior depth appears benign. Because the breasts are dense, physical exam is proportionately more important. A follow Graphical representation of the findings if present may only be seen on the mammography system report. The patient was notified of the results. Froedtert Memorial Lutheran Hospital letter sent: Mammography Benign/Unchanged BI#93A BILATERAL DIAGNOSTIC MAMMOGRAM: 11/6/2002 CLINICAL: Hx of Lt breast focal asymmetric density. Patient has been on Estrogen since 1994. Paternal grandmother dx breast cancer. Comparison is made to exam dated: 8/24/2000 Froedtert Memorial Lutheran Hospital. The tissue of both breasts is heterogeneously dense. This may lower the sensitivity of mammography. A benign appearing calcification is present in the left breast. There is a skin lesion associated with the left breast. A radiopaque marker marks a skin lesion seen in the patient's left breast. Because the breasts are dense, physical exam is proportionately more important. The overall architecture and distribution of the remaining fibroglandular tissue is unchanged. There is a benign focal asymmetric density with an obscured margin in the left breast at 1 o'clock in the posterior depth that most likely represents fibroglandular tissue. Compared to previous films this focal asymmetric density is not significantly changed. No significant masses calcifications or other findings are seen in the right breast. There has been no significant interval change. The Diagonsis is Benign. The Birads Score is 0. The History of Breast Cancer is NoHxBreastCA. The Age of Patient is 45 - 50. The Family History of the Patient is Minor. The History of Surgery is NoSurgery.The Growth of the Tumor is Benign.

BENIGN MAMMOGRAM The 1.2 cm nodular density in the right breast at 12 o'clock in the anterior depth appears benign. Because the breasts are dense physical exam is proportionately more important. There is no mammographic evidence of malignancy. A 1-year screening mammogram is recommended. Graphical representation of the findings if present may only be seen on the mammography system report. The patient was notified of the results. Froedtert Memorial Lutheran Hospital letter sent: Mammography Normal BI#42 A. BILATERAL SCREENING MAMMOGRAM: 3/4/2003 CLINICAL: Screening Self-Referral. Patient has been on Estrogen since 1994. Family history of breast cancer Comparison is made to exam dated :3/20/2001 Froedtert Memorial Lutheran Hospital. There are scattered fibroglandular elements in both breasts that could obscure a lesion on mammography. Scattered benign appearing calcifications are present in both breasts. There are bilateral skin calcifications. Biopsy inferior tissue is present anteriorly in both breasts. No significant masses calcifications or other findings are seen in either breast. There has been no significant interval change. The Diagonsis is Benign. The Birads Score is 1. The History of Breast Cancer is NoHxBreastCA. The Age of Patient is The Age is 61 - 64. The Family History of the Patient is None. The History of Surgery is NoSurgery. The Growth of the Tumor is Benign.

Fig. 3. Examples for real report (upper) and simulated one (lower) for a benign mammogram finding.

The Stanford subset includes only four BIRADS categories - 1,2,5 and 6, for which there is a higher confidence associated with the lesion type (malignant or benign). Each report in both subsets has two separate sections: the findings section that records the imaging observations (Fig. 3) and the impression section that records the overall conclusion/diagnosis. The latter is the clinicians interpretation given the imaging findings. The observations in the findings section should be consistent with the overall impression reported in the impression section.

V. Results

A. Qualitative Analysis

A masked radiologist reviewed and classified 30 reports (Reports evaluation subsection) that were randomly selected for qualitative evaluation. Among these, 0.857 of the real reports were classified correctly by the radiologist and 0.75 of our simulated reports were also classified as real. These results support the high quality of the simulated reports, as most of them were classified by the radiologist as being real. While 0.75 of the RNN-simulated reports were considered by the radiologist to be real reports, only 0.25 of the simulated reports generated by the MRF model (we compared with) were classified as real. These results demonstrate the superiority of this LSTM-based technique over the alternative MRF approach.

B. Quantitative Results

*1) **Style Statistics of the Simulated Reports**:* We analyzed the multi-level style characteristics of the simulated reports and compared them to the equivalent characteristics of the real reports. It can be seen in Table I that the style statistics are similar in both real and our simulated reports. Wilcoxon paired test showed that the similarity between the real and the simulated reports of these characteristics is not significant ($p > 0.05$). In addition, we compared the most frequent keywords each report. The 27 most frequent keywords were selected for each report. Among those, 24 keywords were identical between the real and the simulated reports (0.889 of the frequent words).

TABLE I
Reports' Style - Multi Level Statistics of Mammography Reports

Parameter	Data type	Value
Word length	MCW - Real	5.72 ± 0.11
	Stanford - Real	6.06 ± 0.24
	RNN - Simulated	5.5 ± 0.31
Words in a sentence	MCW - Real	12.2 ± 1.09
	Stanford - Real	16.2 ± 1.48
	RNN - Simulated	14.8 ± 2.34
Words in a report	MCW - Real	235.6 ± 48.9
	Stanford - Real	231.2 ± 134.6
	RNN - Simulated	225.8 ± 66.5

*2) **Contribution of Simulated Reports as Data Augmentation Tool**:* Table II demonstrates the significance of the simulated reports as a data augmentation tool. It shows the contribution of adding simulated data to a limited-sized collection of annotated real data by using 50 real Stanford reports and a varied number of simulated reports (0, 30, 50, 100). The BN was then used to classify the remaining (real) MCW and Stanford reports into benign and malignant classes. Fig. 4 goes even further. It shows the contribution of the additional simulated dataset to the benign and the malignant reports classification, *separately*.

Taking a *fixed* number of training examples for the BN classifier and changing only the percentage of the real and the simulated reports within the training examples, shows that the simulated reports supply a comparable BN classification accuracy, comparing to train it with real reports only (0.943

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.

TABLE II
CLASSIFICATION ACCURACY (GINI COEFFICIENT, *$p < 0.01$, **$p < 0.05$)

Testing	Training Structure	**Our Method**	MRF
MCW **	50 Real Stanford reports only	0.90 ± 0.08	0.84 ± 0.07
MCW *	30 simulated and 50 Real Stanford	0.91 ± 0.02	0.86 ± 0.04
MCW *	50 simulated and 50 Real Stanford	0.94 ± 0.02	0.89 ± 0.03
MCW *	100 simulated and 50 Real Stanford	0.95 ± 0.01	0.89 ± 0.04
Stanford **	50 Real Stanford reports only	0.92 ± 0.03	0.88 ± 0.07
Stanford **	30 simulated and 50 Real Stanford	0.95 ± 0.03	0.91 ± 0.06
Stanford *	50 simulated and 50 Real Stanford	0.98 ± 0.02	0.94 ± 0.05
Stanford *	100 simulated and 50 Real Stanford	0.98 ± 0.01	0.93 ± 0.06

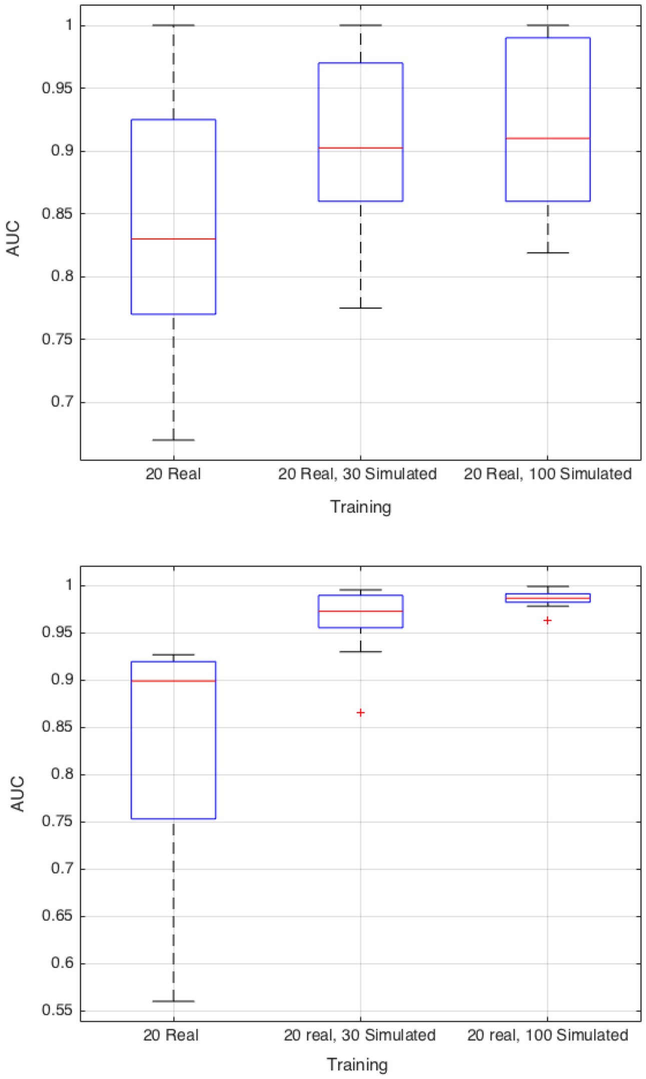


Fig. 4. The effect of using different amounts of simulated data on the accuracy of real report classification (benign(top) /malignant (bottom)).

Vs. 0.935, $p > 0.05$). Means, the simulated reports are a decent replacement to the real ones.

C. Comparison of RNN-LSTM and MRF Generative Models

Table II also shows a comparison between our RNN-LSTM architecture supplied and the MRF. A significant improvement was obtained by using our method ($p < 0.05$, $p < 0.01$, Wilcoxon), in addition to having more robust classification over different datasets.

VI. Discussion

In this work, we developed a method for generating simulated clinical reports, to be used as a data augmentation tool for training machine learning algorithms. In case that the training set is de-identified, we have another added value - simulating reports without risking patients' privacy. We explored the effect of leveraging this simulated data to improve training of a machine learning classifier over the original reports by augmenting the original training set. The results showed that the additional simulated data significantly enhances the end classification accuracy. We evaluated our method in several ways. *First*, a qualitative analysis showed that our simulated reports were realistic in terms of both content and style, since 75% of them were judged by a radiologist to be real reports. The fact that only 25% of the simulated reports generated by the MRF were classified as real by a radiologist supports this conclusion, and demonstrates the superiority of our LSTM-based architecture for this type of text-to-text generation task when both content and style are of utmost importance. *Second*, a quantitative analysis showed that adding a varying number of simulated reports to a training set for learning a machine learning classifier is beneficial; using our LSTM architecture, it is clear that adding larger numbers of simulated reports increased the classification accuracy of new test cases, suggesting that generating simulated reports could add value in training machine learning models. Statistically, our results demonstrate that the LSTM-based method consistently and significantly outperforms the MRF-based method (Wilcoxon).

A limitation of this study is the relatively low diversity of data that was available for our work. Future work will include several directions. We will analyze a other cohorts, including cases from additional different institutions, which would capture more diverse data statistics. We will also explore the training of a word vector embedding model on RadLex, a Radiology-specific ontology that could supply better results if used for training word embeddings. Finally, it will be also interesting to investigate the ability of the method in simulating other types of reports besides mammography; we intend to explore both MRI and CT clinical reports, as their substantial intra-class and inter-class variability make these much more challenging targets for text-to-text generation when compared with highly structured mammography reports.

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.

VII. Conclusion

There is a paucity of labeled training data in the medical domain, and we have showed in the mammography domain that using an LSTM-based technique for simulating text reports can improve performance of machine learning methods trained on datasets comprised of those reports. Our approach was proved to be a useful data augmentation technique when a large amount of labeled data is not available. Towards this end, we have specifically proposed and validated a novel technique that enables generating text-to-text simulations and preserves the content-style characteristics of the original text. To our knowledge, this kind of text-to-text generation framework has not been yet developed, nor has the benefit of using it as a data augmentation strategy been previously demonstrated. According to the presented results, we are confident that our method can have an added value in improving the classification accuracy of a variety of machine learning models that require large amounts of clinical text as their major input.

References

[1] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 901–909.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015, *arXiv:abs/1409.0473*.

[3] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. 21st Int. Conf. Data Eng.*, 2005, pp. 217–228.

[4] E. Bertino, B. Chin Ooi, Y. Yang, and R. Deng, "Privacy and ownership preserving of outsourced medical data," in *Proc. IEEE 21st Int. Conf. Data Eng. (ICDE)* 2005, pp. 521–532.

[5] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:abs/1406.1078*.

[6] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "Rmsprop and equilibrated adaptive learning rates for non-convex optimization," 2015, *arXiv:abs/1502.04390*.

[7] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo, "I2T2I: Learning text to image synthesis with textual data augmentation," 2017, *arXiv:abs/1703.06676*.

[8] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. 21st IEEE Int. Conf. Data Eng.*, 2005, pp. 205–216.

[9] Y. Goldberg and O. Levy, "word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:abs/1402.3722*.

[10] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:abs/1308.0850*.

[11] A. Gupta and P. Mannem, "From image annotation to image description," 2012, in T. Huang, Z. Zeng, C. Li, and C. S. Leung (Eds.), *Neural Information Processing. ICONIP*. Lecture Notes in Computer Science, vol 7667. Berlin, Heidelberg: Springer, pp. 196–204.

[12] D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks," 2016, *arXiv:abs/1609.09106*.

[13] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 279–288.

[14] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:abs/1602.02410*.

[15] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," 2014, *arXiv:abs/1412.2306*.

[16] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2006, pp. 217–228.

[17] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[18] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2005, pp. 49–60.

[19] S. M Thomas, B. Mamlin, G. Schadow, and C. Mcdonald, "A successful technique for removing names in pathology reports using an augmented search and replace method," in *Proc. / AMIA Annu. Symp. AMIA Symp.*, 2002, pp. 777–81.

[20] F. Mairesse and S. Young, "Stochastic language generation in dialogue using factored language models," *Comput. Linguistics*, vol. 40, pp. 763–799, 2014.

[21] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," *Comput. Linguistic*, 2013, pp. 1–12 arXiv:1301.3781.

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119, *arXiv:abs/1310.4546*.

[23] A. H. Oh and A. I. Rudnicky, "Stochastic language generation for spoken dialogue systems," in *Proc. ANLP/NAACL Workshop Conversational Syst.*, vol. 3, 2000, pp. 27–32.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," *Proc. 2014 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, vol. 14, 2014, pp. 1532–1543.

[25] S. Semeniuta, A. Severyn, and E. Barth, "A hybrid convolutional variational autoencoder for text generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 627–637.

[26] T. Sibanda and O. Uzuner, "Role of local context in automatic deidentification of ungrammatical, fragmented text," in *Proc. Main Conf. Human Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2006, pp. 65–73.

[27] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 966–973.

[28] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, 2014, *arXiv:abs/1409.3215*.

[29] L. Sweeney, "k-anonymity: A model for protecting privacy," *IEEE Secur. Privacy*, vol. 10, no. 5, pp. 1–14, Oct. 2002.

[30] G. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," *J. Amer. Medical Informat. Assoc.*, vol. 14, pp. 574–80, 2007.

[31] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," In *Proc. IEEE Conf. Comput. Vision Pattern Recognit*, 2014, pp. 3156–3164, *arXiv:abs/1411.4555*.

[32] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. Very Large Data Base*, 2006, pp. 139–150.

[33] Y. Yang, C. Teo, H. Daum III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2011, pp. 444–454.

[34] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. Zhu, "I2t: Image parsing to text description," *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

[35] X. Zhang and M. Lapata, "Chinese poetry generation with recurrent neural networks," 2014, pp. 670–680.

[36] X. Zhang, F. Yin, Y. Zhang, C. Liu, and Y. Bengio, "Drawing and recognizing chinese characters with recurrent neural network," 2016, *arXiv:abs/1606.06539*.

[37] S. Zhong, Z. Yang, and R. N. Wright, "Privacy-enhancing k-anonymization of customer data," in *Proc. Principles Database Syst.*, 2005, pp. 139–147.

Authorized licensed use limited to: Stanford University. Downloaded on October 13,2020 at 16:46:45 UTC from IEEE Xplore. Restrictions apply.