Generating Mammography Reports from Multi-view Mammograms with BERT

Alexander Yalunin
Sberbank AI Lab
aayalunin@sberbank.ru

Elena Sokolova
Sberbank AI Lab
e.v.sockolova@gmail.com

Ilya Burenko
Sberbank AI Lab
burenko.i.m@sberbank.ru

Alexander Ponomarchuk
Sberbank AI Lab
ponomarchuk@gmail.com

Olga Puchkova
Ilyinsky hospital
o.s.puchkova@icloud.com

Dmitriy Umerenkov
Sberbank AI Lab
d.umerenkov@gmail.com

Abstract

Writing mammography reports can be error-prone and time-consuming for radiologists. In this paper we propose a method to generate mammography reports given four images, corresponding to the four views used in screening mammography. To the best of our knowledge our work represents the first attempt to generate the mammography report using deep-learning. We propose an encoder-decoder model that includes an EfficientNet-based encoder and a Transformer-based decoder. We demonstrate that the Transformer-based attention mechanism can combine visual and semantic information to localize salient regions on the input mammograms and generate a visually interpretable report. The conducted experiments, including an evaluation by a certified radiologist, show the effectiveness of the proposed method. Our code is available at https://github.com/sberbank-ai-lab/mammo2text.

1 Introduction

Breast cancer represents a global healthcare problem (Glo, 2016). Increasing numbers of new cases and deaths are observed in both developed and less developed countries, only partially attributable to the increasing population age. Serial screening with mammography is the most effective method to detect early stage disease and decrease mortality. The goal of screening is to detect breast cancers when still curable to decrease breast cancer-specific mortality (Duffy et al., 2020). The European Society of Breast Imaging (EUSOBI) together with 30 national breast radiology bodies recommend that only qualified radiologists should be involved in screening programs. (Sardanelli et al., 2017).

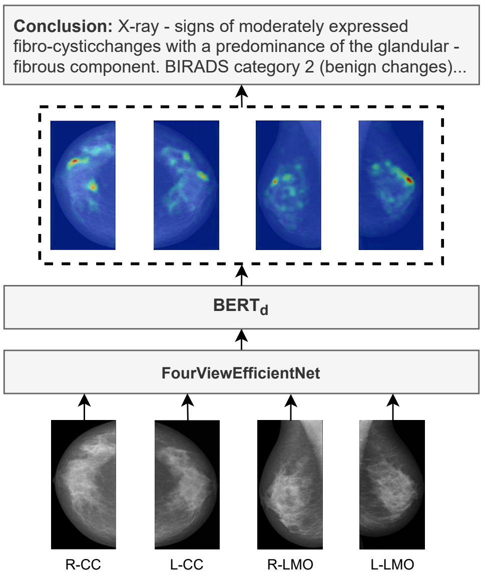

Conclusion: X-ray - signs of moderately expressed fibro-cysticchanges with a predominance of the glandular - fibrous component. BIRADS category 2 (benign changes)...
BERT$_d$
FourViewEfficientNet
R-CC
L-CC
R-LMO
L-LMO

Figure 1: Overview of the proposed framework for interpretable mammography report generation. For examples of generated reports, see appendix.

As the amount of organized breast screening programs grows across the world, the burden on radiologists increases with it. In National screening programs such as in Holland or Sweden, radiologists may need to read 100 radiology images per hour (Abbey et al., 2020). With a growing number of screening programs, we need more trained radiologists and new technologies that can make their workflow more effective. Since one of the most time consuming procedures in radiology is writing medical-imaging reports, we explore the potential for deep-learning to automatically generate diagnostic reports of screening mammograms.

The rapid evolution of deep learning and artificial intelligence technologies enables them to be used as a strong tool for providing clinical decision-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 153–162
November 7–11, 2021. ©2021 Association for Computational Linguistics

making support to the medical community. While many problems in the area of medical imaging and text analysis have been addressed effectively, there is no known approach to generating clinical reports for mammography studies. There are various reasons for this, such as the requirements regarding the accuracy, completeness and diagnostic relevance of the clinical information contained in the report.

In this article, we present a framework (Figure 1) that takes mammograms as an input, automatically generates mammography reports, and visualizes the attention of the model to provide the interpretability of the process.

We use an encoder-decoder architecture, where the encoder extracts visual features and the decoder generates reports. We adopt a convolutional neural network, specifically EfficientNet (M Tan, 2019), to extract visual features of the four images, corresponding to the four views used in screening mammography. For language modeling, we utilize BERT (Devlin et al., 2018), inserting an additional attention sub-layer to perform multi-head attention over the regional feature embeddings produced by the encoder. We modify the Transformer-based (Vaswani et al., 2017) attention mechanism such that it attends to the visual information on four mammography views and previously generated words. We use the attention scores to build visually interpretable image-text attention mappings.

In addition to that, we conduct a series of in-depth quantitative and qualitative experiments with the help of an experienced radiologist to demonstrate the clinical validity of our approach. We compare the predictions of our models with the ground truth to understand where the models make mistakes and demonstrate that our best model successfully describes different parts of the breast, and detects pathological regions and abnormalities. We evaluate the image-text attention mappings to demonstrate the interpretability of our model.

As far as we are aware, our work represents the first attempt to generate the mammography report using deep-learning.

To summarize, we make the following contributions in this paper:

- We propose a novel framework for mammography report generation using EfficientNet in the encoder and BERT in the decoder.

- We demonstrate that the Transformer-based attention mechanism can combine visual and textual information to localize salient regions on the input mammograms and generate a visually interpretable report.

- We conduct doctor evaluation and extensive experiments with automatic metrics to show the effectiveness of the proposed framework.

- We conduct a qualitative analysis including interpretation of image-text attention mappings to demonstrate how the model is able to generate mammography reports in a meaningful way.

2 Related work

The task of image captioning is creating a model that given a previously unseen query image generates a caption that is both grammatically and semantically correct. The main approaches to image captioning are retrieval-based, template-based and novel caption generation. In retrieval-based methods (Hodosh et al., 2013), (Ordonez et al., 2011) candidate captions for query images are selected from a pool of existing captions based on some measure of similarity. The downside of this approach is the inability to generate novel image-specific captions. In template-based methods (Farhadi et al., 2010), (Kulkarni et al., 2013), (Li et al., 2011) image captions are generated by filling the blanks in fixed templates. These methods can generate grammatically and semantically correct novel captions not present in the training set but cannot generate variable-length captions. Novel caption generation methods (Xu et al., 2015), (Yao et al., 2017), (You et al., 2016) use a representation of the query image as an input for a language model responsible for generating the captions. This approach follows the encoder-decoder architecture first applied to machine translation tasks (Cho et al., 2014).

To generate an image caption, a representation of the image must first be constructed either via generating handcrafted features or extracting such features automatically, for example using deep neural networks. Examples of hand-crafted features are local binary patterns (Ojala et al., 2002), scale-invariant keypoints (Lowe, 2004), or histograms of oriented gradients (Dalal and Triggs, 2005). Automatic feature extraction from images is commonly used by applying convolutional neural networks (CNN) (LeCun et al., 1998) to the query image. These features may be further enhanced, for example by using a spatial Transformer (Pedersoli et al., 2017).

A sub-field of image captioning is diagnostic captioning (DC). Diagnostic captioning is automatic generation of diagnostic text based on a set of medical images of a patient. DC systems can increase the speed of producing a report for experienced physicians and decrease the number of diagnostic errors for inexperienced doctors (for a recent survey on DC methods see (Pavlopoulos et al., 2021)). The majority of the work in DC is done using encoder-decoder architecture. In addition to evaluation of grammatical and semantical correctness of captions, which is commonly assessed by calculating lexical overlap between generated captions and ground truth (Pavlopoulos et al., 2019), DC quality can be assessed by clinical correctness by conducting clinical experiments with physicians evaluating the generated reports (Zhang et al., 2019), (Liu et al., 2019).

Language models commonly used in DC usually apply recurrent neural networks (RNN) such as LSTM (Hochreiter and Schmidhuber, 1997), see (Vinyals et al., 2015) (Xu et al., 2015), with works using Transformer-based models beginning to appear (Chen et al., 2020).

A common approach in DC is the use of 'visual attention' that allows the decoder to focus on particular areas of input images when generating the captions (Jing et al., 2017), (Yuan et al., 2019). Such mechanisms also can be used to highlight the regions of interest on the input images adding to the interpretability of the models (Zhang et al., 2017).

3 Data

The dataset is based on data from a breast screening program in one of the Russian regions. The dataset includes about 25K screening mammography studies with clinical reports. All exams include four standard mammography views: R-CC (right craniocaudal), L-CC (left craniocaudal), R-MLO (right mediolateral oblique), L-MLO (left mediolateral oblique), with image height and width of 4644 by 3510 pixels respectively. Each study contains a brief text conclusion, clinical report and BI-RADS class. Mammography reports are written in Russian, examples in this article are translated into English. On average, the mammography report contains 55 words. All personally identifiable information has been deleted by the clinics.

We split the dataset into the training, validation and test subsets in the proportion of 91%, 4% and 5% respectively (having 22463, 934 and 1229 cases in each subset). The splits are the same for encoder pretraining and for the text generation model.

№	Target	Cases	
		Train	Val
0	Lesions	2936	147
1	Shadows	1339	71
2	Calcifications	1108	61
3	Fibrosis	12441	649
4	Skin Thickening	106	6
5	BI-RADS > 1	18919	997
6	BI-RADS > 2	2153	114

Table 1: Binary targets extracted from mammography reports for encoder pre-training.

4 Method

We start with describing the formal definition of the task. Given four mammogram images S we try to generate a sequence of words Y that represents the mammography report:

$$S = \{I_{LCC}, I_{RCC}, I_{LMLO}, I_{RMLO}\}$$

$$Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \mathbf{y}_i \in R^K$$

where $I_\star$ represents an image of one of the four projections, K is the size of the vocabulary and C is the length of the generated report. Given a set of images and the corresponding mammography report Y, the model maximizes the negative conditional log-likelihood:

$$\theta^* = \arg\max_\theta \sum_{(S,Y)} \log p(Y \mid S; \theta)$$

where θ is the parameters of the model. The chain rule then allows the log-likelihood of the joint probability to be factored as the sum of individual conditionals:

$$\log p\left(y_{1:C} \mid S; \theta\right) = \sum_{i=1}^{C} \log p\left(y_i \mid S, y_{1:i-1}; \theta\right)$$

The model we introduce is fundamentally an encoder-decoder. The encoder receives the set of projections as input and extracts the set of visual features using a convolutional neural network. Next, the Transformer-based decoder generates the complete mammography report given the set visual features of the images.

4.1 Encoder Pretraining

We use a deep multi-view (N Wu, 2019) CNN based on EfficientNet B0 (M Tan, 2019). We chose EfficientNet B0 because it is relatively lightweight and fits in GPU memory when using high resolution images. We have one EfficientNet instance for all views (R-CC, L-CC, R-MLO, L-MLO), i.e. model weights are shared. The first convolutional layer is replaced to accept a one-channel image. The last fully-connected layer of EfficientNet is discarded. Outputs from all four views are averaged by channels and one fully connected layer is added.

The encoder is pretrained to predict multilabel targets important for diagnosis in mammography screening, shown in Table 1. The binary targets were extracted with regular expressions from text descriptions of the studies. Targets № 0-4 are typical pathological changes in breasts tissues. During training, the images are cropped and resized to 1350x900 px.

4.2 Encoder Fine-tuning

Given a set of images S, FourViewEfficientNet (FVEN) extracts a set of visual features:

$$X = \{\mathbf{x}_1, \dots, \mathbf{x}_r\} = \mathrm{FVEN}(S), \mathbf{x}_i \in R^d$$

where r is the number of sub-regions and d is the embedding size of the sub-region. Similarly to (Xu et al., 2015) we extract feature maps from the last convolutional layer, which yields a 4 × 43 × 29 × 1280 tensor. The dimensions of this tensor are equal to the number of images, height, width and the number of channels respectively. The number of sub-regions $r = 4988$ (reshaped from 4 × 43 × 29). Each sub-region as an output of the last convolution layer is represented as an m-dimensional vector, where m is equal to the number of channels of the last convolutional layer, here $m = 1280$. They are then passed through a linear layer with a ReLU activation and the output size $d = 768$.

4.3 Decoder

For the decoder part we use BERT (Devlin et al., 2018) with an additional attention sub-layer. At this point, we could use a more natural Transformer-based decoder architecture like GPT (Radford et al., 2019), but as shown in (Rothe et al., 2020) in the encoder-decoder architectures BERT as the decoder performs better than GPT. BERT uses masked language modeling for pretraining bidirectional word representations and provides contextualized word representations during the fine-tuning stage.

To use BERT as the decoder we need to insert an additional attention sub-layer, which performs multi-head attention over the output of the encoder, i.e. regional visual features. To emphasize this change we denote our decoder model as $\mathrm{BERT_d}$. The predicted sequence of words can be obtained by:

$$y_i = \mathrm{BERT_d}(\mathrm{X}, \mathbf{y}_1, \dots, \mathbf{y}_{\mathrm{i}-1})$$

In our experiments we compare two variants of BERT. The first variant is RuBERT (Kuratov and Arkhipov, 2019): a BERT pretrained on the general corpus of Russian texts. The second is BERT pretrained exclusively on a medical corpus. We omit the pretraining details as they are beyond the scope of this article.

4.4 Attention mechanism

We now briefly describe how the attention mechanism is implemented in the Transformer (Vaswani et al., 2017). The input consists of three parts: queries Q, keys K and values V. The output is computed as:

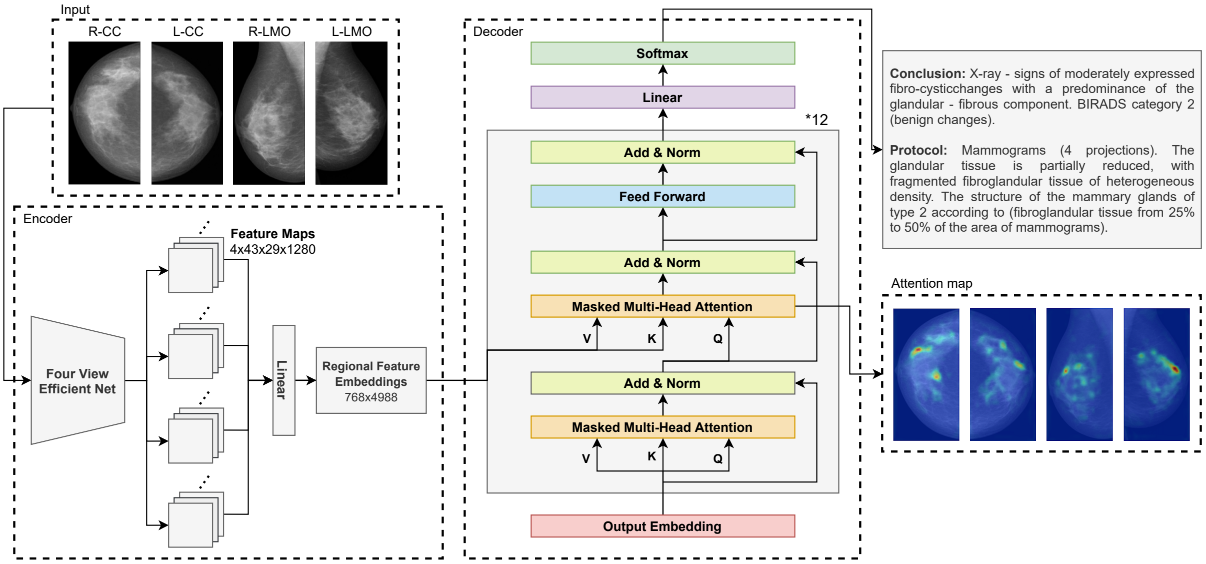
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_{attn}}}\right)V$$

The matrices Q, K and V are computed as follows:

$$Q = Q_{in} \cdot W_Q, K = K_{in} \cdot W_K, V = V_{in} \cdot W_V$$

where $W_Q, W_K, W_V \in R^{d_{model} \times d_{attn}}$ are the embedding matrices, d_{model} is the dimensionality of the input and output, and d_h is the dimensionality of one head. This procedure is repeated h times, where h is the number of heads, which produces h different sets of queries, keys and values.

Each decoder layer consists of two sub-layers which employ this multi-head attention mechanism, but differ in the inputs Q_{in}, K_{in} and V_{in}. The self-attention in the first sub-layer can attend only to the outputs of the previous decoder layer, in this case $Q_{in} = K_{in} = V_{in}$. In the second sub-layer the attention mechanism attends to both the outputs of the encoder X and the outputs of the previous sub-layer Z, thus: $K_{in} = V_{in} = X$ and $Q_{in} = Z$. Recall that the outputs of the encoder are regional feature embeddings of the input image set. This



Conclusion: X-ray - signs of moderately expressed fibro-cysticchanges with a predominance of the glandular - fibrous component. BIRADS category 2 (benign changes).

Protocol: Mammograms (4 projections). The glandular tissue is partially reduced, with fragmented fibroglandular tissue of heterogeneous density. The structure of the mammary glands of type 2 according to (fibroglandular tissue from 25% to 50% of the area of mammograms).

Figure 2: The architecture of our encoder-decoder model. FourViewEfficientNet in the encoder takes four views and produces the feature maps tensor which is then passed through a linear layer. The resulting matrix is used as value and key matrices in the attention mechanism in the decoder layers. The decoder produces the mammography report and the image-text attention mappings.

way of using the Transformer attention mechanism allows for each word in the generated output sequence to attend over all regions of the input image set S, which leads to the possibility of building interpretable image-text attention mappings.

5 Experiments

A series of retrospective data experiments were carried out to evaluate the performance of the developed models. First, we measure the performance of our models with the commonly used natural language generation metrics (NLG), including CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002). We compare four model variants with a random baseline, where the predicted report is a real report for a different patient. Then, we evaluate the text reports generated by our model with the help of an experienced radiologist, both quantitatively and qualitatively. We provide a comprehensive description of the experimental procedure together with the obtained results in this and the following section.

5.1 Model Variants

In this subsection we describe different model variants. All hyperparameters and configurations in the following models are the same, except for the changes described below.

- **FEN2RND** An EfficientNet pretrained on the ImageNet dataset (Deng et al., 2009) and used four times in the FourViewEfficientNet, paired with randomly initialized BERT. The encoder returns only one embedding of all four views.
- **FEN2RND+att** Same as FEN2RND , but the encoder outputs embeddings for each sub-region and the decoder attention mechanism is applied over these embeddings. The same attention mechanism is used in the following models as well. This novelty aims to demonstrate the effect of multi-head attention over regional image information.
- **MFEN2RUBERT** A FourViewEfficientNet additionally trained to classify mammogramm images paired with RuBERT: a BERT pretrained on the corpus of Russian texts. This baseline aims to demonstrate the effect of using pretrained models.
- **MFEN2MBERT** A FourViewEfficientNet additionally trained to classify mammogramm images paired with BERT pretrained exclusively on a medical corpus.

5.2 Implementation details

An important difference between the model variants is the way the encoder extracts visual features. In the FEN2RND the encoder outputs one 768-dimensional vector which we feed into the encoder. In the model variants that use an image-text

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
Random	0.463	0.394	0.343	0.308	0.462	0.299	0.225
FEN2RND	0.540	0.488	0.449	0.418	0.534	0.320	0.952
FEN2RND+att	0.552	0.503	0.466	0.435	0.549	0.329	0.935
MFEN2RUBERT	**0.594**	**0.533**	**0.485**	**0.446**	**0.575**	**0.340**	0.883
MFEN2MBERT	0.572	0.514	0.471	0.437	0.548	0.331	**0.954**

Table 2: Quantitative evaluation of model variants on the validation dataset - includes only automated metrics.

	Automated evaluation							**Doctor evaluation**		
Model	B1	B2	B3	B4	R	M	C	CAL	LES	Rating
Random	0.549	0.464	0.400	0.349	0.472	0.285	0.628	0.114	0.302	1.000
FEN2RND	0.626	0.571	0.529	0.495	0.569	0.359	1.382	0.116	0.359	2.810
FEN2RND+att	0.641	0.590	0.550	0.517	0.583	0.371	1.440	0.143	0.531	4.439
MFEN2RUBERT	**0.662**	**0.608**	**0.567**	**0.534**	**0.590**	**0.374**	**1.649**	0.263	**0.642**	5.585
MFEN2MBERT	0.646	0.590	0.548	0.515	0.587	0.364	1.559	**0.270**	0.630	**5.887**

Table 3: The comparison between automated metrics and doctor evaluation on the doctor dataset. B{n} denotes BLEU using up to n-grams. R, M, C denote ROUGE-L, METEOR and CIDEr, respectively. On doctor evaluation CAL denotes Calcifications and LES denotes Lesions.

attention mechanism the encoder outputs 4 × 43 × 29 × 1280 feature maps which are then flattened and linearly transformed into a 4988 × 768 tensor.

We used the default BERT configurations with 12 layers, 12 heads and the dimensions of all hidden states and word embeddings equal to 768. The models are trained under softmax cross entropy loss with Adam optimizer (Kingma and Ba, 2014) and half precision. We used linear learning rate decay with 5e-5 initial learning rate. All models were trained for 5 epochs with batch size equal to 4. At generation step we used beam size equal to 5.

The maximum length of the generated report C was set to 224. The vocabulary size K of the RuBERT tokenizer is equal to 120,000 and the vocabulary size of BERT trained on a medical corpus is equal to 40,000.

We use the encoder-decoder architecture, the trainer pipeline and the language model implementations from HuggingFace library (Wolf et al., 2020). We modify the encoder-decoder logic so that the image model can be used as the encoder.

Each model was trained for 1 day on one NVIDIA Tesla V100 GPU.

5.3 Doctor Evaluation

To assess the efficiency of the proposed models, we conduct an experiment involving a board-certified radiologist with sixteen years of experience in the writing and evaluation of mammography diagnostic reports. For the experiment, an extra set of data was prepared consisting of 150 anonymized breast X-rays with clinical reports. The doctor is asked to evaluate six reports for each case: the ground truth, four reports that came from model variants and a random report for another case. For the doctor evaluation we use the two most important predetermined clinical criteria: Calcifications and Lesions. These criteria have been selected for evaluation as the most critical for the correct diagnosis. Each criterion has been classified by the doctor as "is in the image but not in the text”; "is in the text, but not in the image"; "is both in the text and in the image"; "is neither in the text nor the image". In addition to that, the doctor gave an overall assessment of each report on a scale of one to ten, based on completeness, relevance and accuracy. We normalize this rating so that the ground truth prediction gets the highest rating and the random prediction gets the lowest. To avoid bias, the reports for each case were given in a randomized order, so that the doctor does not have information on the source of any individual report within each study.

6 Results

6.1 Quantitative Analysis

6.1.1 Report Generation

The report generation performance is measured on two datasets. Table 2 presents the results on the validation dataset using NLG metrics only. Here the metrics were measured for each BI-RADS separatingly and then the average was taken. Table 3 compares side-by-side the automated metrics and doctor evaluations on the dataset made for doctor evaluation described in Section 5.3.

We make the following observations: 1) The use of the attention mechanism demonstratings a significant improvement in the performance of the model. The model FEN2RND+att that introduces attention demonstrates improvement in doctor rating from 2.81 to 4.44, as well as an improvement in all NLG metrics. This demonstrates the effectiveness of the proposed visual-text attention mechanism. 2) The second significant improvement comes from the use of pretrained models on the general domain in both encoder and decoder of MFEN2RUBERT . This model variant demonstrates the best performance on automated metrics among all model variants. Calcifications and Lesions improved as well, while doctor rate rose from 4.4 to 5.5. 3) MFEN2MBERT is our best performing model according to human evaluation. Surprisingly it does not show the best performance on automated metrics. After a qualitative examination in Section 6.2 it becomes clear that the model pretrained on the medical domain employs medical terms like calcifications, shadows, and lesions more accurately than the model pretrained only on the general domain. It is a common known fact that the automated metrics do not measure aspects relevant to the specific domain.

6.1.2 Classification

In order to validate our results shown in Tables 2 and 3 we conduct an additional experiment with the output from BERT. As mentioned in Section 4.1 we are able to mine a binary vector of length 5 for each of the five classes (see Table 1). We use this script to parse BERT's output and a vector of binary variables. This approach allows us to compare classification metrics of BERT and the pretrained multilabel classification encoder (Section 4.1). We compare Matthews Correlation Coefficient (Chicco and Jurman, 2020) for each of five binary targets between labels mined from text generated by BERT, labels predicted by the pretrained encoder, and labels from a random doctor's report from the validation dataset.

We see that for targets such as lesions, shadows and skin thickening BERT is able to improve classification results while for such targets as Calcifications and Fibrosis BERT degrades the encoder's results. We argue that the high level convolutional features that BERT utilizes within its attention mechanism (see Figure 2) allow the generative model to capture spatial information that leads to substantially better results in classification of skin thickening than compared to plain convolutional models such as multi-label classification FVEN.

Target	BERT	Encoder	Random mean±std
Lesions	**0.449**	0.417	0.002±0.031
Shadows	**0.411**	0.394	0.001±0.03
Calcifications	0.363	**0.379**	0.003±0.029
Fibrosis	0.294	**0.341**	0.001±0.021
Thick skin	**0.615**	0.417	0.0±0.0

Table 4: MCC score

6.2 Qualitative Analysis

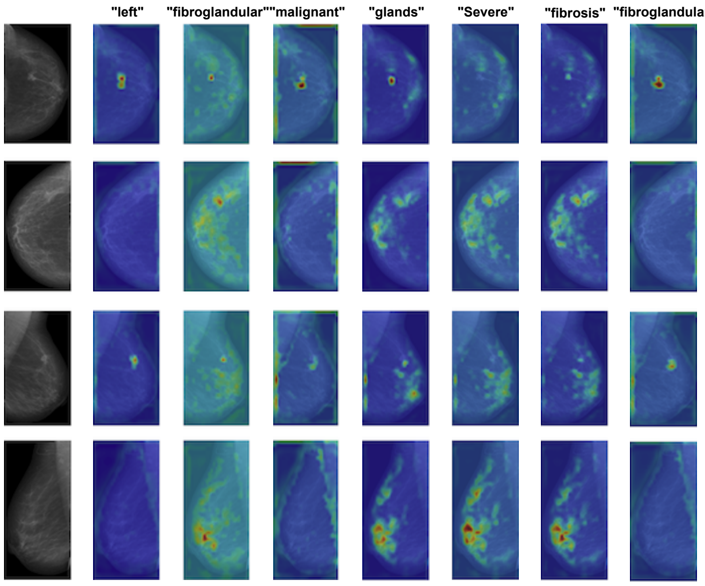
6.2.1 Case Study

Along with the described quantitative experiments to assess the quality of the developed models together with the expert, we perform an extensive clinical analysis of generated reports on a subset of cases. Here we analyze three cases where we compare mammography reports generated by FEN2RND and MFEN2MBERT models with the ground truth report. Due to space constraints, we could not show the examples and direct the reader to the appendix. The first case is shown in Figure 4, the second and the third cases are shown in Figure 5.

In every case MFEN2MBERT not only correctly predicts the breast density but also accurately identifies pathological regions. Some of the cases where the location of the lesion is described imprecisely could be explained by the presence of bordering regions. The same terms are used for describing the site of abnormality. Different doctors have different descriptions for normal and abnormal, which makes the generated text sequence diverse.

Unlike MFEN2MBERT , FEN2RND fails to identify abnormalities in all three cases, although it predicts breast density fairly well. Sometimes the skin and the nipple are also not describe correctly. This is important because in some cases only these regions of the mammogram are indicative of breast cancer in patients, and would lead the radiologist to recommend additional examination.

In the first case (Figure 4) MFEN2MBERT describes the nipple, but does not see its retraction. One of the reasons for this could be a rare occurrence of this symptom in the training set, so with more data the model could identify this as well as



Ground Truth: Conclusion: X-ray signs of malignant lesion of the left breast. X-rays are signs of a weakly expressed fibro-cystic changes with a predominance of the fibrous component. BIRADS category 5 (findings indicate breast cancer). Protocol: Mammograms (4 projections). The mammary glands are symmetrical. The nipples are not retracted. The skin in the upper quadrants of the left breast is deformed. Premammary spaces are unchanged. On a fatty background residual fibroglandular tissue and severe fibrosis along the ducts. The structure of the mammary glands of type 1 according to (fibroglandular tissue less than 25% of the mammogram area). In the upper outer quadrant of the left breast is an asymmetric area of with indistinct spiculated contours, measuring 1, 9x1, 7 cm.

MFEN2MBERT: Conclusion: X-ray signs of malignant lesion of the ***left*** breast. BIRADS category 6 (verified ***malignant*** process). Protocol: Mammograms (4 projections). The mammary ***glands*** are symmetrical. The nipples are not retracted. The skin is not thickened. Premammary spaces are unchanged. Adipose tissue predominates. Upper - outer quadrants are partially dence due to residual heterogeneous ***fibroglandular*** tissue. ***Severe*** fibrosis along the ducts. Diffuse ***fibrosis***. The structure of the mammary glands of type 1 according to (***fibroglandular*** tissue less than 25% of the mammogram area). In the upper outer qadrant of the left breast, near the pectoral muscle there is a dense lesion with irregular shape, 2, 3x2, 3x2, 3cm. the contours are indistinct.

Figure 3: Visualization of image-text attention mappings from MFEN2MBERT between four mammography views and generated report.

it identifies the presence of lesions.

In the second case (Figure 5) MFEN2MBERT describes the abnormality and reports the shape of the lesion, which is crucial as cancer and benign lesions have different shapes.

In the first and second cases MFEN2MBERT correctly classifies BI-RADS, unlike FEN2RND . However, in the first case it predicts BI-RADS-3 instead of 4, which could be the result of a mistake by the model or caused by lesions which feature signs that border on benign and malignant, such as fibroadenoma and mucinous cancer. If the problem is caused by borderline signs, then future work could explore using more data for training the model on this special subtype of lesion.

6.2.2 Interpret Model Attention

In order to interpret the output of our model, we visualize the image-text attention mappings from our best model MFEN2MBERT between four mammography views and the generated report. Together with a doctor, we analyze them for the presence or absence of clinical correlation between the generated report and the regions of the mammogram that the model pays attention to. We analyze three cases. The first case in shown in Figure 3; the second (Figure 6) and third (Figure 7) cases can be found in appendix.

For the first case (Figure 3) the model successfully detects the area (“upper outer quadrant of the left breast”) which is abnormal (“dense lesion”). Thus, the model detects and describes a malignant lesion, which is a good result that may lead to a high PPV in screening.

In the second case (Figure 6) several right correlations between the text and the mammogram areas can be seen. First, the model is looking directly at fibroglandular tissue and does not classify it as an abnormality. Therefore, the model can predict breast density well, which is very important, since breast density is associated with an increased risk of developing breast cancer and requires additional examination, such as breast ultrasound or MRI. Secondly, no abnormalities are present either in the image or in the report from the model. This is likewise very important as it may lead to a low false positive rate and a low callback rate – metrics of breast screening programs.

In the third case (Figure 7) the model does not work correctly. It describes the fibroglandular tissue subtype while looking at the subcutaneous fat. The density type is also incorrectly specified.

7 Conclusion

In this paper we present a first-of-its-kind framework for generating mammography reports given four mammography views using deep-learning. Our model utilizes pretrained models including EfficientNet for visual extraction and BERT for report generation. We demostrate that the Transformer-based attention mechanism that simultaneously attends to four mammography views and text from the report significantly improves the performance.

Our method provides a novel perspective for breast screening: generating mammography reports and providing image-text attention mappings, which makes the automatic breast screening process semantically and visually interpretable. The validity of our approach is confirmed by the corresponding doctor evaluation. In the conducted qualitative analysis we demonstrate that our best model successfully detects pathological regions, and describes abnormalities and parts of the breast.

References

2016. Globocan online analysis. *http://globocan.iarc.fr/Pages/burden$_s$el.aspx*.

Craig K Abbey, Michael A Webster, Tanya Geertse, Danielle van der Waal, Eric Tetteroo, Ruud Pijnappel, Mireille JM Broeders, and Ioannis Sechopoulos. 2020. Sequential reading effects in dutch screening mammography. In *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, volume 11316, page 113160G. International Society for Optics and Photonics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Stephen W Duffy, László Tabár, Amy Ming-Fang Yen, Peter B Dean, Robert A Smith, Håkan Jonsson, Sven Törnberg, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, et al. 2020. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*, 126(13):2971–2979.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Q Le M Tan. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *PMLR*.

J Park Y Shen Z Huang N Wu, J Phang. 2019. Deep neural networks improve radiologists' performance in breast cancer screening. *arXiv preprint arXiv:1903.08297*.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36.

John Pavlopoulos, Vasiliki Kougia, Ion Androutsopoulos, and Dimitris Papamichail. 2021. Diagnostic captioning: A survey. *arXiv preprint arXiv:2101.07299*.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE international conference on computer vision*, pages 1242–1250.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Francesco Sardanelli, Hildegunn S Aase, Marina Álvarez, Edward Azavedo, Henk J Baarslag, Corinne Balleyguier, Pascal A Baltzer, Vanesa Beslagic, Ulrich Bick, Dragana Bogdanovic-Stojanovic, et al. 2017. Position paper on screening for breast cancer by the european society of breast imaging (eusobi) and 30 national breast radiology bodies from austria, belgium, bosnia and herzegovina, bulgaria, croatia, czech republic, denmark, estonia, finland, france, germany, greece, hungary, iceland, ireland, italy, israel, lithuania, moldova, the netherlands, norway, poland, portugal, romania, serbia, slovakia, spain, sweden, switzerland and turkey. *European radiology*, 27(7):2737–2743.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436.