Informatics in Medicine Unlocked 24 (2021) 100557

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Automated radiology report generation using conditioned transformers

Omar Alfarghaly [a,*], Rana Khaled [b], Abeer Elkorany [a], Maha Helal [b], Aly Fahmy [a]

[a] Computers and Artificial Intelligence, Cairo University, Cairo, Egypt
[b] National Institute of Cancer, Cairo University, Cairo, Egypt

ARTICLE INFO

Keywords:
Report generation
Transformers
GPT2
Transfer learning
X-ray
Deep learning

ABSTRACT

Radiology report writing in hospitals is a time-consuming task that also requires experience from the involved radiologists. This paper proposes a deep learning model to automatically generate radiology reports given a chest x-ray image from the public IU-Xray dataset. Our work consists of three stages: (1) Fine-tune a pre-trained Chexnet to predict specific tags from the image. (2) Calculate weighted semantic features from the predicted tag's pre-trained embeddings. (3) Condition a pre-trained GPT2 model on the visual and semantic features to generate the full medical reports. We analyze the generated reports using word-overlap metrics while also adding new meaningful semantic-based similarity metrics. The proposed model, which we call CDGPT2, surpassed most non-hierarchical recurrent models and transformer-based models in quantitative metrics while being considerably faster to train. Moreover, the model does not require a specific vocabulary and can be trained on different datasets without changing the architecture. Furthermore, we include a qualitative analysis from a radiologist from Egypt's national institute of cancer which showed that 61.6% of the generated reports on the test set were expertly written, and only 10% contained false information. We represent the first work to condition a pre-trained transformer on visual and semantic features to generate medical reports and to include semantic similarity metrics in the quantitative analysis of the generated reports.

1. Introduction

Medical imaging techniques are widely used in hospitals worldwide. The detailed information generated from medical images is necessary for diagnosing illnesses or tracking patients' progress. However, every image requires a radiologist to carefully examine and write a full-text report to describe the findings. Diagnosing medical images requires an appropriate amount of experience from the radiologists to develop more confident and accurate reports. Nevertheless, many reports conclude with indecisive findings that require the patient to take further tests, including pathology or other advanced imaging methods, as the spectrum of possible cases is too broad. Furthermore, a more glaring issue is the amount of time it takes the radiologist to write a full-text report. It would take on average 10 min or more based on the radiologist's degree of experience, so this would prove very time-consuming when considering the number of cases a radiologist should investigate per day, and in crowded hospitals, regions, and cities, this would be problematic. These reasons combined provided good motives for us to research deep learning models capable of automating report writing.

Some researchers have studied the automatic report generation problem [1–3]. In their works, they depend on convolution-recurrent architectures (CNN-RNN) introduced in image captioning research such as [4,5] and visual attention on the recurrent decoder like in Ref. [6]. Recently, the Natural Language Processing (NLP) community has been shifting from recurrent models as attention-only-based models known as transformers [7–9] proved faster to train and can leverage GPU parallelization instead of the sequential nature of the recurrent models. With the wide use of transformer-based models, the NLP community entered the transfer learning phase, which the computer vision entered after Imagenet [10]. Most NLP research now achieves superior results by fine-tuning an existing pre-trained transformer model on a huge corpus. This motivated us to investigate using existing pre-trained transformers with generative capabilities like GPT2 [9] while conditioning them on the visual features and semantic tags embeddings, as conditioning a pre-trained transformer provides benefits like faster training, removes the need to specify a vocabulary, and already learned word-structure and punctuation.

In this paper, we present a conditioned transformer-based model that

* Corresponding author. Giza Governorate, 12613, Egypt.
E-mail addresses: o.mohamed@grad.fci-cu.edu.eg (O. Alfarghaly), r_hkhaled@hotmail.com (R. Khaled), a.korani@fci-cu.edu.eg (A. Elkorany), dr.mahahelal@yahoo.com (M. Helal), a.fahmy@fci-cu.edu.eg (A. Fahmy).

https://doi.org/10.1016/j.imu.2021.100557
Received 21 December 2020; Received in revised form 23 February 2021; Accepted 20 March 2021
Available online 26 March 2021

we call *CDGPT2* for 'conditioned distil generative pre-trained transformer 2', which generates a full report given a chest X-Ray image. We trained the model using the Indiana University chest X-Ray dataset (IU-XRay) [11], which is publicly available. The model consists of an encoder that outputs visual and semantic features from the image and a decoder to generate the words. Similar to Ref. [3], the encoder is a pre-trained Chexnet model [12], that was fine-tuned to predict multiple tags from the image. The predicted scores of each tag are then multiplied by their corresponding pre-trained word2vec embeddings [13], which were trained on a vast medical corpus, and are passed with the visual features to the decoder. The decoder is a pre-trained distilGPT2 [14] that is being conditioned on the visual features and the tags' embeddings to produce the full-text report. Then we compare our quantitative results with previous works, while adding new meaningful quantitative metrics that focus on semantic similarity, and found that our model outperformed most non-hierarchical recurrent models and transformer-based models in the average Bleu scores. We also include a qualitative analysis from a radiologist from Egypt's national institute of cancer.

To summarize, our contributions are as follows:

- Used a pre-trained transformer to remove problems like vocabulary selection and punctuation handling.
- Introduced a new conditioning technique for the model CDGPT2 that proved faster to train and outperformed most non-hierarchical recurrent models and previous transformer-based models in the average Bleu scores.
- Added semantic similarity metrics besides the word-overlap metrics to enhance the quantitative analysis.

The rest of the paper is organized as follows: Section 2 reviews the related works, Section 3 describes the model architecture and methods, Section 4 presents and compares the results, and Section 5 concludes the paper.

2. Related work

Image Captioning: Image captioning is the problem of generating text to describe the input image. The problem gained popularity with the rise of deep learning techniques, as many works adopted the CNN-RNN architecture [4]. Following the success of the attention concept introduced in Ref. [15], more works started adding visual attention to their CNN-RNN architecture like in Refs. [6,16]. Moreover, works like [5] added semantic attention with the visual attention. Furthermore, hierarchical recurrent models were introduced to solve the problems that occur when generating long captions like in Ref. [17]. Few papers attempted to use transformer-based models [8,9] as decoders in the image captioning domain, although they are faster to train and produce state-of-the-art results in most NLP problems. One of the papers that attempted to condition a GPT2 model was [18] by adding new key and value weights in the self-attention module to be projected into the decoder's self-attention space.

Medical Image Captioning: There have been several works attempting to generate medical reports from their corresponding images. The first of which to use a CNN-RNN approach was [19] to predict tags used to form a structured report for chest X-ray images from the IU-Xray dataset. The first work to use attention on the medical image was [2], long-short-term-memory cells (LSTM) [20] were used to produce a report of five sentences on a private dataset of pathology images. A framework to generate natural reports for the Chest-Xray14 dataset [21], using private reports, was introduced in Ref. [22], using a non-hierarchical CNN-LSTM architecture and attention on semantic and visual features. Natural reports on the IU-Xray dataset were generated in Ref. [1] by getting visual features from a VGG network pre-trained on Imagenet and introducing the concept of co-attention which is a combined attention mechanism on both the visual and predicted tags embeddings. The co-attention output is then passed to a hierarchical LSTM, one for sentences and one for words, to generate the reports. The multi-view information of the IU-Xray dataset was leveraged in Ref. [3] by getting the visual features and tags' prediction from the front and side images of the patient using a Resnet152 trained on the Chexpert dataset [23] then using hierarchical LSTMs like that of [1] to generate the reports. Knowledge graphs with prior knowledge on chest findings were utilized in Ref. [24], the graph node features are extracted from the IU-Xray images using Chexnet models [12], and hierarchical LSTMs, with attention over the graph, were used to generate the full report. A cross-modal retrieval method was used in Ref. [25] to retrieve abnormal findings given images from the IU-Xray dataset by learning visual-semantic embeddings on the images and the reports, which are then used to measure the similarity between a new image and the existing findings to retrieve the closest report. A feature pyramid network was used in Ref. [26] to concatenate features from different models, pre-trained on ImageNet, given chest X-ray images to detect labels that depict the image; these labels are used to create a vector that describes the major findings in the image. The semantic distance between the image's vector and the encoded reports in the IU-Xray dataset was used to retrieve the most relevant report and remove parts from the report with no evidence. Few works used transformers as decoders in radiology report generation like [27], which used a custom transformer as the decoder with a bottom-up region detector and a top-down visual encoder to generate reports for the IU-Xray dataset. A custom transformer was also used in Ref. [28] with an extra relational memory unit to generate reports for the IU-Xray dataset. The front and side chest images are passed through a visual extractor to get a sequence of visual features, from pre-trained models like VGG and Resnet, that will be passed to the encoder and decoder to generate the reports.

We will also be conditioning on visual and semantic features. However, we used a pre-trained GPT2 as the decoder and adopted a similar conditioning technique to that of [18], but we condition on visual and weighted semantic features.

3. Methodology

As shown in Fig. 1, the model architecture consists of three major components, the visual model, semantic features' generation, and the decoder. The visual model, acting as the encoder, is responsible for predicting the tags associated with the image and generating the visual features. The semantic features are computed by multiplying the tags' confidence scores with their corresponding pre-trained embeddings, as explained in Section 3.2. The decoder is a pre-trained transformer-based model conditioned on the visual and semantic features.

3.1. Visual features

The image is first passed through the visual model to produce the visual features and the tags' predictions. Our base model is a Chexnet[1] [12], which is a Densenet121 model [29] pre-trained on ChestX-ray14 dataset [21] to detect and localize 14 types of diseases or anomalies from the images. While the base model can provide good visual features, we found that 14 tags were not enough to provide good and varied semantic features, so we fine-tuned the model to classify the manual tags from the IU-Xray dataset by removing the final layer and adding a new final layer containing 105 nodes for the most occurring manual tags from the dataset.

$$L(b) = -\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i_t} \,.\, log\,\widehat{y}_{i_t} + \left(1-y_{i_t}\right) .\, log\left(1-\widehat{y}_{i_t}\right) \quad (1)$$

[1] We used the pre-trained model available here: https://Github.com/brucechou1983/Chexnet-Keras.

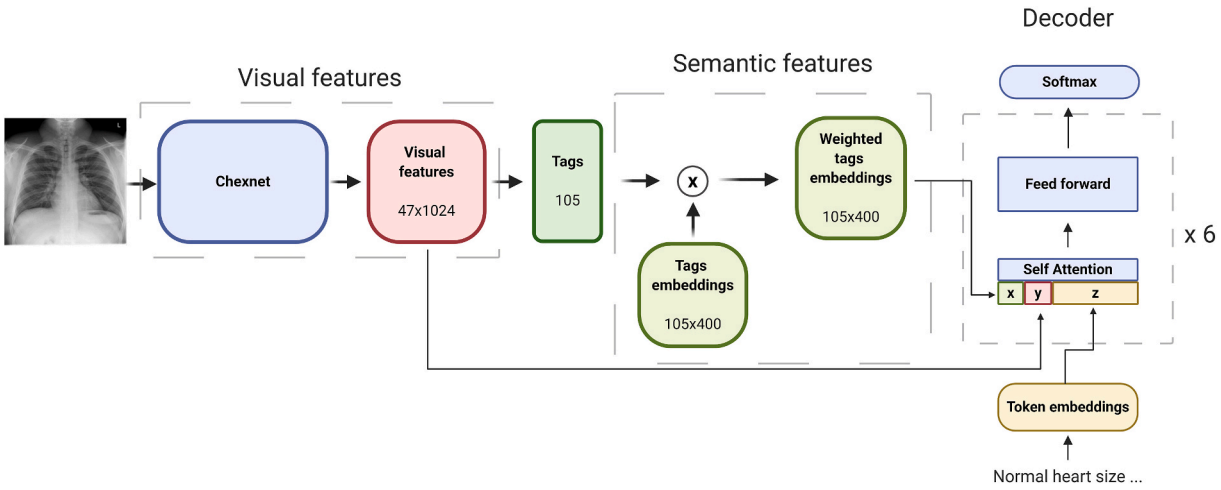


Fig. 1. Proposed CDGPT2 model architecture. *Tags* represent a vector of size 105 containing the independent probabilities of each tag between 0 and 1. *Tags embeddings* is a matrix of pre-trained word2vec embeddings.

As shown in Equation (1), the problem was treated as multi-label classification with binary cross-entropy loss where $L(b)$ is the loss for batch b, T denotes the number of tags, and N denotes the batch size. The model was trained end-to-end using mini-batch gradient descent, 32 images per batch, and Adam optimizer [30]. All the parameters in the model were left to be fine-tuned.

3.2. Semantic features

The visual model outputs a vector of size 105, representing each tag's independent confidence score between 0 and 1. To condition the decoder on the semantic features along with the visual features, a pre-trained word2vec embeddings were used for the tags trained on biomedical texts from MEDLINE/PubMed[2] which were introduced in Ref. [31].[3] If a tag consists of more than one word, the embedding of that tag will be the average of its individual words' embeddings.

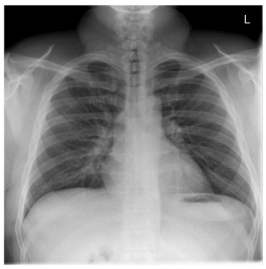
The tags' embeddings is a 105×400 matrix. The final semantic features are the multiplication of the tags' scores with their corresponding embeddings. The weighted tags' embeddings help the decoder learn which semantics are relevant to the image and which are not.

3.3. Decoder

A pre-trained distilGPT2 [14] was used as the decoder. distilGPT2 is a compressed version of GPT2 that consists of 6 layers, a hidden layer size of 768, 12 heads, and 82 million overall parameters. distilGPT2 is twice as fast as a normal GPT2 model. The model was pre-trained on OpenWebTextCorpus,[4] a reproduction of OpenAI's WebText dataset which was used to train the original GPT2 model [9]. The output layer consists of 50257 nodes representing the English language's byte-pair encoding. The same output layer was kept as it can generate all the medical terms.

3.4. Conditioning details

The standard transformer model contains a self-attention module consisting of keys, queries, and values, as shown in Equation (3) where *softmax* is defined as in Equation (2), Z represents the input token embeddings, and W_q, W_k, W_v represent the weights for the queries, keys, and values respectively.



Impression: Negative chest x-XXXX.

Findings: Cardiac and mediastinal contours are within normal limits. The lungs are clear. Bony structures are intact.

Manual Tags: Normal

Fig. 2. A sample image from the IU-Chest X-ray dataset. The report consists of an *Impression* which serves as a title, *Findings* which contain the full report, and a *Manual Tags* section listing some keywords to describe the image.

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{2}$$

$$SA(Z) = softmax\left((ZW_q)(ZW_k)^T\right)(ZW_v) \tag{3}$$

The conditioning proposed in Ref. [18] adds new keys and values parameters, initialized randomly, to project the encoder output to the decoder's attention space. We adopted a similar technique while also adding semantic features with the visual as shown in Equation (4), where *CSA* is the conditioned self-attention, Z represents the input token embeddings, X is the semantic features, Y is the visual features, and U_k, H_k, U_v, and H_v represent the new keys and values added for the semantic and visual features respectively.

$$CSA(X,Y,Z) = softmax\left((ZW_q)\begin{bmatrix} XU_k \\ YH_k \\ ZW_k \end{bmatrix}^T\right)\begin{bmatrix} XU_v \\ YH_v \\ ZW_v \end{bmatrix} \tag{4}$$

4. Experiments

This section will describe the dataset, explain the implementation details, describe the baselines that were tested against, and analyze the quantitative and qualitative results.

4.1. Dataset

The Indiana University chest X-ray dataset [11] consists of 7430 images of the front and side chest X-rays, belonging to 3825 patients and their corresponding reports. The report, as shown in Fig. 2, consists of an

[2] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.
[3] https://www.gitmemory.com/issue/RaRe-Technologies/gensim-data/28/496097622.
[4] https://skylion007.Github.io/OpenWebTextCorpus/.

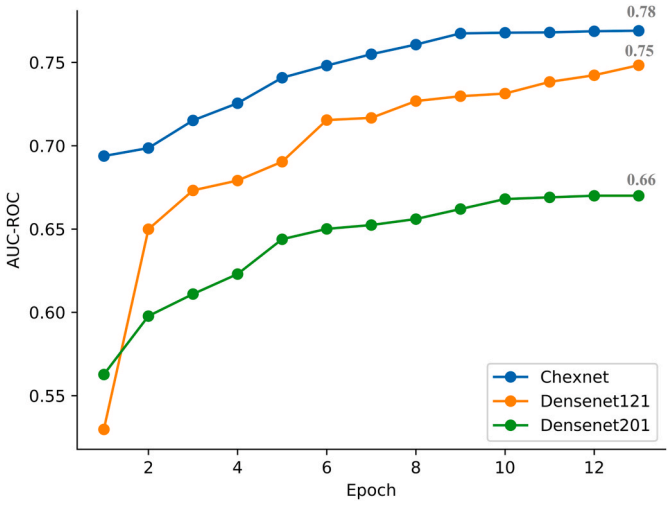


Fig. 3. Different trials to predict the 105 tags from the X-Ray images. The AUC-ROC represents the average AUC-ROC on the test set.

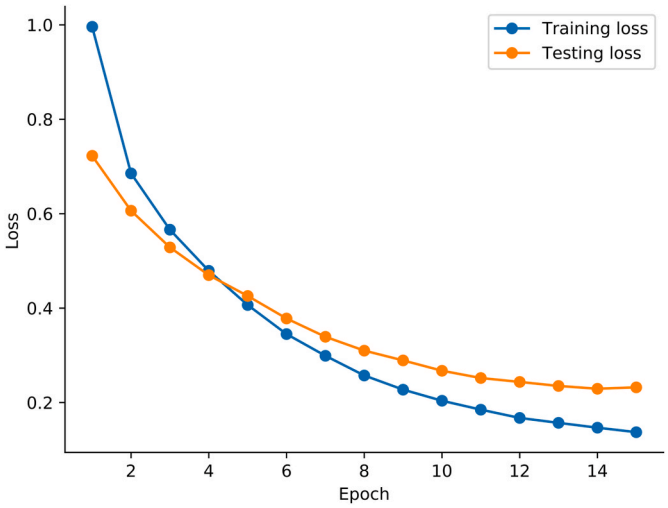


Fig. 4. The training and testing loss of the CDGPT2 model.

Impression, which serves as a title or summary of the report, *Findings* which include the report in detail, and finally, there are *Manual Tags* associated with each report. There are two types of tags in this dataset: (1) MTI tags that were extracted from the MTI indexer[5] from the reports. (2) Manual tags from MESH[6] and RadLex[7] that were manually associated by trained coders to describe each case.

The concatenation of *Impression* and *Findings* was used as the target report as in Ref. [1]. Moreover, we used the manual tags as the targets to train the visual model. There were 187 unique tags; the tags that appeared less than 25 times were removed, which resulted in 105 tags. We also removed the confidential information from the reports which were set to XXXX in the dataset. Furthermore, 500 images were selected randomly as the test set.

4.2. Implementation details

Visual Model: For the multi-label tags classification, the images were resized to be 224 × 224, as the base Chexnet model was trained on 224 × 224 images. The images were resized using interpolation and anti-aliasing, and normalized by subtracting from the mean and dividing by the standard deviation. All the parameters in the pre-trained Chexnet were left trainable during the fine-tuning. A batch size of 32, Adam optimizer [30], a learning rate of 1e-3 decaying to 1e-7, and class weights to handle imbalance between the tags were used. The weight for each class is defined as shown in Equation (5) where *CW* represents the class weight, c_i is the input class, P_{c_i} the number of positive instances of class c_i, and N the number of samples in the training set.

$$CW(c_i \in \{0,1\}) = \begin{cases} \frac{P_{c_i}}{N} \ if \ c_i \ is \ 0 \\ \frac{N - P_{c_i}}{N} \ if \ c_i \ is \ 1 \end{cases} \quad (5)$$

GPT2 Decoder: The visual model's weights and the tags' pre-trained embeddings are frozen during the training of the GPT2 model. A dropout layer [32] with a drop probability of 0.4, chosen experimentally, was added to the visual and semantic features before sending them to the decoder. Adam optimizer was used with a constant learning rate of 1e-3. Furthermore, a max sequence length of 200 tokens was used during training and inference. *startseq* was used as a beginning of sentence token, the standard GPT2 end of sentence $< |endoftext| >$ token was kept, and'<' was used as the pad token. Fig. 4 shows the convergence of the training and testing sets. In inference, a beam search of width 7 was selected.

Python 3.7 and Tensorflow 2.1 were used in our experiments. Moreover, Huggingface's transformers [33] was used as the GPT2 base. The code is available publicly on Github.[8]

4.3. Visual model results

Since it is a multi-label classification problem, the average area under the receiver operating characteristics curve (AUC-ROC) was used as a metric to compare the different fine-tuning trials, as shown in Fig. 3. The results show that a model pre-trained on chest X-ray images like Chexnet outperforms other models that were pre-trained on ImageNet.

4.4. Captioning baselines

We compare our method against four types of models. The first type is non-hierarchical recurrent models like [4,5,16]. The second type is hierarchical recurrent models with attention like [1,3]. The third type is retrieval-based models like [25,26]. The fourth type is transformer-based models like [27,28]. Moreover, to better showcase the added benefit of combining visual and semantic features to condition the decoder, we compare against our model when only using visual or semantic features as proposed in Ref. [18]. Additionally, to compare the training speed of using a transformer model, we implemented a recurrent model with visual and semantic attention, shown in Fig. 5, which we call *VSGRU* for visual and semantic gated recurrent units. The same visual model was used to get the visual features and the tags predictions for VSGRU, and a threshold of 0.1 was used to decide if a tag exists. The existing tags' pre-trained embeddings are averaged and concatenated to the visual features to create a vector of 1424 elements. Bahdanau's attention [15] was used on the features, and a gated recurrent unit (GRU) [34] cell was used to output the words. Furthermore, a vocabulary of the most occurring 1000 words in the training set was used.

4.5. Quantitative analysis

We used natural language generation evaluation (NLG) [35] to provide us with quantitative metrics, shown in Table 1, for the generated

[5] https://ii.nlm.nih.gov/MTI/.
[6] https://www.nlm.nih.gov/mesh/meshhome.html.
[7] http://www.radlex.org/.

[8] Our code repository: https://Github.com/omar-mohamed/GPT2-Chest-X-Ray-Report-Generation.

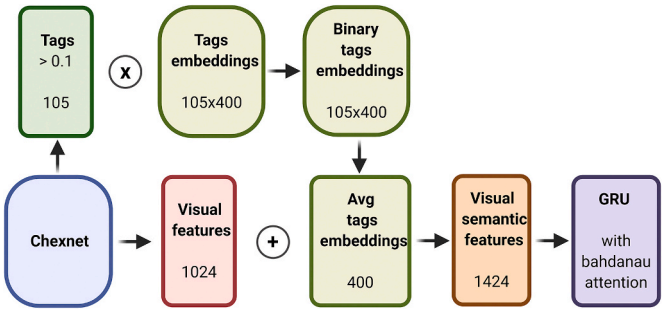


Fig. 5. A figure showing our custom visual-semantic-recurrent model (VSGRU).

reports. We provide word-overlap based metrics using Bleu [36], METEOR [37], ROUGLE-L [38], and Cider [39]. However, we believe that these metrics are not enough to evaluate sensitive systems when it comes to captioning, as discussed in Refs. [35,40], especially when there is only one reference sentence like in the IU-Xray dataset. To overcome this challenge, other quantitative methods that focus on semantic similarity were used, as shown in Table 2, like skip-thought cosine similarity [41], embedding average, which is the cosine similarity between the average of the words' embeddings, vector extrema [42], and greedy matching [43]. To our knowledge, we are the first work to incorporate semantic similarity in the quantitative analysis of radiology report generation.

The results show that in word-overlap metrics, our model CDGPT2 scored a higher average Bleu score than most non-hierarchical models (first section in Table 1). RNCM [19] has a higher bleu average as they scored a high Bleu-1 score, but they only predict tags to generate a limited controlled report, which explains the drop in Bleu-3 and Bleu-4 using this method. In the transformer-based section (fourth section in Table 1), CDGPT2 outperformed [27] in all metrics except Cider. However, it failed to outperform [28] as it leverages information from both the front and side chest views before outputting the report compared to our model, which works on single images. Furthermore, CDGPT2 did not manage to surpass hierarchical recurrent models (second section in Table 1) and [25] from cross-modal retrieval methods (third section in Table 1). Moreover, our custom recurrent model *VSGRU* managed to outperform all other methods in the Cider score. Still, overall, word-overlap metrics alone might not be representative of the quality of the reports.

In the semantic similarity metrics shown in Table 2, we compared our method against our custom recurrent model *VSGRU*. The *CDGPT2* model showed improved results in all semantic similarity metrics, which showcase its ability to produce more semantically relevant reports. Moreover, we compared our methods when using only visual or semantic features and found that combining the visual and semantic features improved both word-overlap and semantic similarity metrics.

4.6. Qualitative analysis

This section provides a qualitative analysis from a radiologist from Egypt's national institute of cancer with five years of experience reading chest X-ray images. The radiologist was given access to the images and the ground-truth reports to evaluate the automatically-generated reports.

The radiologist classified the generated reports into accurate, missing details, and false reports. The accurate reports are the ones that include most of the vital information and contain no false information; these reports could pass as reports written by experts. The missing details section contains the reports with no false information but missed some important details provided by real-world experts. The false reports are reports that included wrong information and incorrect overall diagnosis. Fig. 6 shows an example for each type of classification performed by the radiologist and the corresponding tags' highlighting on the image. The highlighting is performed using Grad-CAM [44].

Table 3 shows the results of the qualitative analysis. On the test set, the model was able to generate accurate reports for 61.7%, 28.2% had missing information, and 10.2% had a false diagnosis. The model was able to generate reports correctly for 99% of the normal cases, with only two false reports. For the abnormal cases, the model could generate 36.5% of the reports correctly, 47.1% lacked sufficient details, and 16.4% had false reports.

The results overall are promising, but there is still a capacity for improvements. The results show that the model is a bit biased towards normal cases, as it is often the case with medical datasets, and often lacks the necessary details to describe the different irregularities present in the image. Nonetheless, the reports classified as missing details by the radiologist also contain useful information that can be used to speed up the report writing process.

4.7. Time performance analysis

We compare the time it takes the *CDGPT2* model to train against the custom recurrent model *VSGRU* in Fig. 7. The same GPU *Tesla P100 PCIe* and CPU *2-core Xeon 2.2* GHz were used in the experiments, and the same optimizer as Adam, learning rate of *1e-3*, and batch size of 16 were also kept. The time taken for the batch loading and pre-processing was neglected to focus on the actual training time. The results show that

Table 1
Quantitative results using word-overlap metrics. CDGPT2 is our method when using both visual and semantic features. CDGPT2-vis-only and CDGPT2-sem-only represent our method when only using visual or semantic features. VSGRU is a custom recurrent model that uses visual and semantic attention.

Method	Word-overlap metrics						
	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
CNN-RNN [4]	0.316	0.211	0.140	0.095	0.159	0.267	0.111
LRCN [16]	0.369	0.229	0.149	0.099	0.155	0.278	0.190
ATT-RK [5]	0.369	0.226	0.151	0.108	0.171	0.323	0.155
RNCM [19]	**0.785**	0.144	0.047	0	–	–	–
(Ours) VSGRU	0.347	0.221	0.156	0.116	0.150	0.251	**0.413**
Co-ATTN [1]	0.517	0.386	0.306	0.247	0.217	0.447	0.327
MvH [3]	0.529	0.372	0.315	0.255	0.343	0.453	–
SentSAT+KG [24]	0.441	0.291	0.203	0.147	–	0.367	0.304
CVSE [25]	0.192	–	–	0.036	0.077	0.153	–
FFL+CFL [26]	0.56	**0.51**	**0.5**	**0.49**	**0.55**	**0.58**	–
RTMIC [27]	0.350	0.234	0.143	0.096	–	–	0.323
RM+MCLN [28]	0.470	0.304	0.219	0.165	0.187	0.371	–
(Ours) CDGPT2-vis-only	0.340	0.209	0.138	0.091	0.153	0.281	0.229
(Ours) CDGPT2-sem-only	0.357	0.224	0.151	0.103	0.149	0.275	0.267
(Ours) CDGPT2	0.387	0.245	0.166	0.111	0.164	0.289	0.257

Table 2
Quantitative results on the test set using semantic-based similarity metrics.

Method	Semantic similarity metrics			
	SkipThoughtCS	Embedding Avg	Vector Extrema	Greedy Matching
(Ours) VSGRU	0.506	0.820	0.455	0.692
(Ours) CDGPT2-vis-only	0.594	0.861	**0.523**	0.701
(Ours) CDGPT2-sem-only	0.630	0.857	0.520	0.684
(Ours) CDGPT2	**0.632**	**0.863**	0.514	**0.715**

training a non-hierarchical recurrent model takes double the time of training a DistilGPT2 as the decoder. Moreover, it is worth noting that using a pre-trained transformer in *CDGPT2* was faster to converge as it reached a Bleu-1 score of 0.38 compared to 0.34 from *VSGRU*. Furthermore, the average inference time to generate a full report for a single image with a beam width of 7 using *CDGPT2* is 4.4 s, compared to 16.8 s using *VSGRU*.

5. Conclusion

This paper introduced a new conditioning technique to condition a pre-trained DistilGPT2 on visual and semantic features to generate full-text reports for chest X-Ray images. Conditioning a pre-trained transformer model proved: (1) faster to train, (2) eliminated the need to specify a vocabulary for the model, (3) provided better word-overlap based quantitative metrics compared to most previous non-hierarchical recurrent models and transformer-based methods. We also represent the first work to add semantic similarity quantitative metrics that work on the embedding level to evaluate medical reports as we believe that standard word-overlap-based methods are not enough for sensitive systems. Moreover, we include a qualitative analysis for our work from a radiologist from Egypt's national institute of cancer that showed promising results.

However, some shortcomings should be addressed for the model to

Table 3
The results of the qualitative analysis performed by a radiologist.

Images	Accurate	Missing details	False
All (500)	308 (61.6%)	141 (28.2%)	51 (10.2%)
Normal (201)	199 (99%)	0 (0%)	2 (1%)
Abnormal (299)	109 (36.5%)	141 (47.1%)	49 (16.4%)

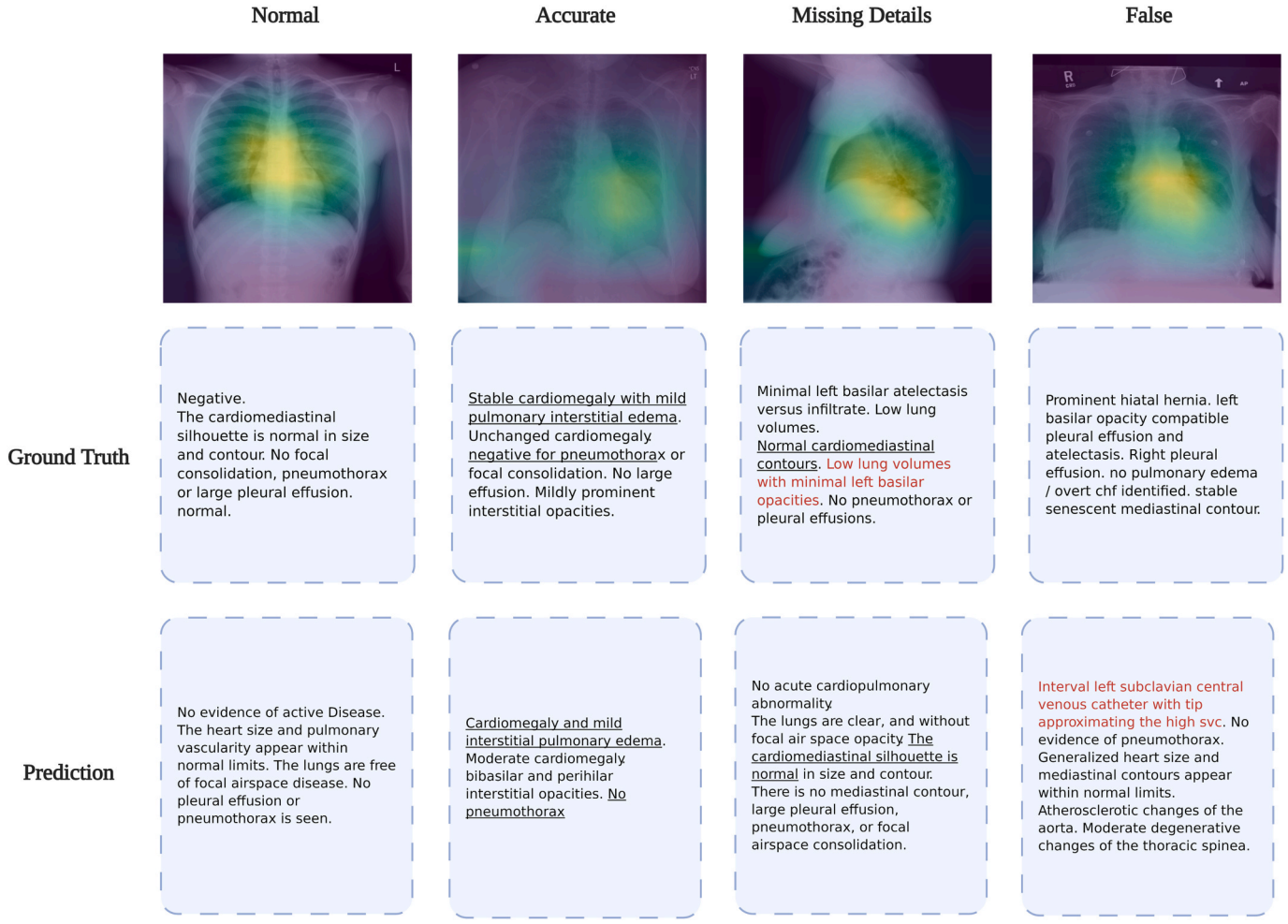


Fig. 6. Example predictions of our CDGPT2 model. The underlined texts are cases in which the model could detect abnormalities and describe them similar to the ground-truth reports. The red text shows wrong or missing information in the generated prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

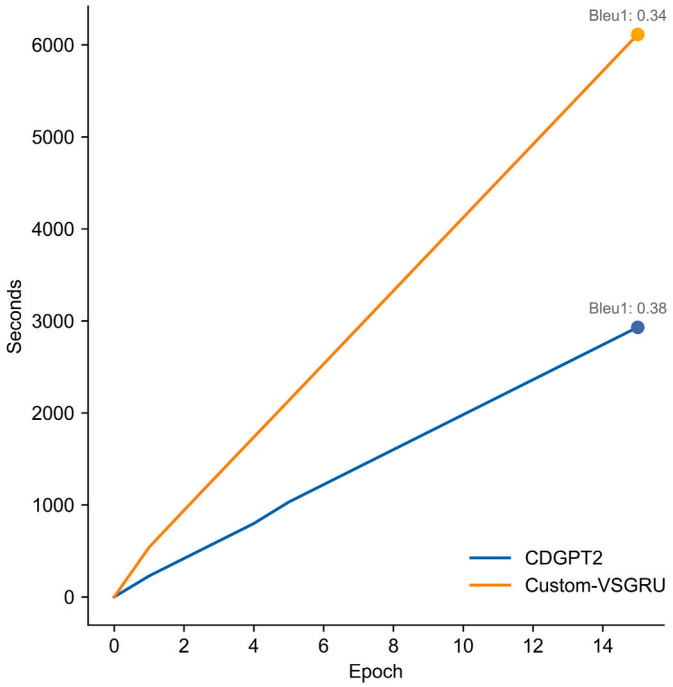


Fig. 7. A figure showing the training time of the custom recurrent model *VSGRU* and the transformer-based model *CDGPT2*.

become more reliable, as many of the generated reports miss some information. We believe that the small size of the dataset limits the model's generalization and makes it prone to over-fitting. Hence, we highly encourage the release of larger datasets that contain multiple reports per image.

In conclusion, we propose a new deep learning model that uses visual and semantic features to condition a pre-trained transformer. We add semantic similarity metrics besides word-overlap metrics for the quantitative analysis. We hope our work will encourage more researchers to consider pre-trained transformers as decoders to generate medical reports to help with the small datasets. We also invite critical thinking in the quantitative methods used to evaluate sensitive systems like medical reports. Finally, we released the code and the trained models publicly to be used or examined by other researchers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None. No funding to declare.

References

[1] Jing B, Xie P, Xing EP. On the automatic generation of medical imaging reports. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th annual Meeting of the Association for computational linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics; 2018. p. 2577–86.

[2] Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: a semantically and visually interpretable medical image diagnosis network. In: 2017 IEEE Conference on computer Vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017. p. 3549–57.

[3] Yuan J, Liao H, Luo R, Luo J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P, Khan A, editors. Medical image Computing and computer assisted intervention - MICCAI 2019 - 22nd international conference, Shenzhen, China, October 13-17, 2019, proceedings, Part VI. Springer; 2019. p. 721–9. vol. 11769 of Lecture Notes in Computer Science.

[4] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society; 2015. p. 3156–64.

[5] You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society; 2016. p. 4651–9.

[6] Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: neural image caption generation with visual attention,. In: Bach FR, Blei DM, editors. Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR.org; 2015. p. 2048–57. vol. 37 of JMLR Workshop and Conference Proceedings.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R, editors. Advances in neural information processing systems 30: annual Conference on neural information processing systems 2017, 4-9 december 2017, Long Beach, CA, USA; 2017. p. 5998–6008.

[8] Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human Language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and short papers). Association for Computational Linguistics; 2019. p. 4171–86.

[9] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. 2019.

[10] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25: 26th annual Conference on neural information processing systems 2012. Proceedings of a meeting held december 3-6, 2012, Lake tahoe, Nevada, United States; 2012. p. 1106–14.

[11] Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani SK, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Medical Informatics Assoc. 2016;23(2):304–10.

[12] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding DY, Bagul A, Langlotz C, Shpanskaya KS, Lungren MP, Ng AY. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. CoRR; 2017. vol. abs/1711.05225.

[13] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems 26: 27th annual Conference on neural information processing systems 2013. Proceedings of a meeting held december 5-8, 2013, Lake tahoe, Nevada, United States; 2013. p. 3111–9.

[14] Sanh V, Debut L, Chaumond J, Wolf T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR; 2019. abs/1910.01108.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y, editors. 3rd international Conference on Learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; 2015.

[16] Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 2017;39(4):677–91.

[17] Krause J, Johnson J, Krishna R, Fei-Fei L. A hierarchical approach for generating descriptive image paragraphs. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017. p. 3337–45.

[18] Ziegler ZM, Melas-Kyriazi L, Gehrmann S, Rush AM. Encoder-agnostic adaptation for conditional language generation. CoRR; 2019. abs/1908.06938.

[19] Shin H, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society; 2016. p. 2497–506.

[20] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.

[21] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017. p. 3462–71.

[22] Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society; 2018. p. 9049–58.

[23] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball RL, Shpanskaya KS, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: The thirty-third AAAI Conference on artificial intelligence, AAAI 2019, the thirty-first innovative Applications of artificial intelligence conference, IAAI 2019, the ninth AAAI Symposium on educational Advances in artificial intelligence, EAAI 2019,

Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press; 2019. p. 590–7.
[24] Zhang Y, Wang X, Xu Z, Yu Q, Yuille AL, Xu D. When radiology report generation meets knowledge graph. In: The thirty-fourth AAAI Conference on artificial intelligence, AAAI 2020, the thirty-second innovative Applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press; 2020. p. 12910–7.
[25] Ni J, Hsu C, Gentili A, McAuley JJ. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. In: Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on empirical Methods in natural language processing: findings, EMNLP 2020, online event, 16-20 november 2020. Association for Computational Linguistics; 2020. p. 1954–60.
[26] Syeda-Mahmood TF, Wong KCL, Gur Y, Wu JT, Jadhav A, Kashyap S, Karargyris A, Pillai A, Sharma A, Syed AB, Boyko OB, Moradi M. Chest x-ray report generation through fine-grained label learning. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical image Computing and computer assisted intervention - MICCAI 2020 - 23rd international conference, Lima, Peru, October 4-8, 2020, proceedings, Part II. Springer; 2020. p. 561–71. vol. 12262 of Lecture Notes in Computer Science.
[27] Xiong Y, Du B, Yan P. Reinforced transformer for medical image captioning. In: Suk H, Liu M, Yan P, Lian C, editors. In machine Learning in medical imaging - 10th international workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, proceedings. Springer; 2019. p. 673–80. vol. 11861 of Lecture Notes in Computer Science.
[28] Chen Z, Song Y, Chang T, Wan X. Generating radiology reports via memory-driven transformer. In: Webber B, Cohn T, He Y, Liu Y, editors. In Proceedings of the 2020 Conference on empirical Methods in natural language processing, EMNLP 2020, online, november 16-20, 2020. Association for Computational Linguistics; 2020. p. 1439–49.
[29] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017. p. 2261–9.
[30] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international Conference on Learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; 2015.
[31] McDonald RT, Brokos G, Androutsopoulos I. Deep relevance ranking using enhanced document-query interactions. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on empirical Methods in natural language processing, Brussels, Belgium, october 31 - november 4, 2018. Association for Computational Linguistics; 2018. p. 1849–60.
[32] Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15 (1):1929–58.
[33] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J. Huggingface's transformers: state-of-the-art natural language processing. CoRR; 2019. abs/1910.03771.
[34] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on deep Learning, December 2014; 2014.
[35] Sharma S, Asri LE, Schulz H, Zumer J. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. CoRR; 2017. abs/1706.09799.
[36] Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual Meeting of the Association for computational linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL; 2002. p. 311–8.
[37] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,. In: Goldstein J, Lavie A, Lin C, Voss CR, editors. Proceedings of the workshop on intrinsic and extrinsic evaluation measures for Machine translation and/or summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005. Association for Computational Linguistics; 2005. p. 65–72.
[38] Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization Branches out, (Barcelona, Spain). Association for Computational Linguistics; July 2004. p. 74–81.
[39] Vedantam R, Zitnick CL, Parikh D, Cider ". Consensus-based image description evaluation. In: IEEE Conference on computer Vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society; 2015. p. 4566–75.
[40] Kougia V, Pavlopoulos J, Androutsopoulos I. A survey on biomedical image captioning. CoRR; 2019. abs/1905.13302.
[41] Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Urtasun R, Torralba A, Fidler S. Skip-thought vectors. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, December 7-12, 2015, Montreal, Quebec, Canada; 2015. p. 3294–302.
[42] Forgues G, Pineau J, Larchevêque J-M, Tremblay R. Bootstrapping dialog systems with word embeddings. In: Nips, modern machine learning and natural language processing workshop, vol. 2; 2014.
[43] Rus V, Lintean MC. An optimal assessment of natural language student input using word-to-word similarity metrics. In: Cerri SA, Clancey WJ, Papadourakis G, Panourgia K, editors. Intelligent tutoring systems - 11th international conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings. Springer; 2012. p. 675–6. vol. 7315 of Lecture Notes in Computer Science.
[44] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128(2):336–59.