Contrastive Pre-training and Linear Interaction Attention-based Transformer for Universal Medical Reports Generation

Zhihong Lin[a,1], Donghao Zhang[b,1], Danli Shi[c], Renjing Xu[d], Qingyi Tao[f], Lin Wu[e], Mingguang He[g], Zongyuan Ge[b,*]

[a]*Faculty of Engineering, Monash University, Clayton, VIC, 3800, Australia*
[b]*Monash eResearch Center, Monash University, Clayton, VIC, 3800, Australia*
[c]*State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangzhou, 510060, China*
[d]*Microelectronics Thrust, The Hong Kong University of Science and Technology (Guangzhou), Nansha, Guangzhou, Guangdong, 511400, China*
[e]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230000, China*
[f]*NVIDIA AI Technology Center, 038988, Singapore*
[g]*Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, 3002, Australia*

Abstract

Interpreting medical images such as chest X-ray images and retina images is an essential step for diagnosing and treating relevant diseases. Proposing automatic and reliable medical report generation systems can reduce the time-consuming workload, improve efficiencies of clinical workflows, and decrease practical variations between different clinical professionals. Many recent approaches based on image-encoder and language-decoder structure have been proposed to tackle this task. However, some technical challenges remain to be solved, including the fusion efficacy between the language and visual cues

*Corresponding author
[1]Indicates the equal contribution.

and the difficulty of obtaining an effective pre-trained image feature extractor for medical-specific tasks. In this work, we proposed the weighted query-key interacting attention module, including both the second-order and first-order interactions. Compared with the conventional scaled dot-product attention, this design generates a strong fusion mechanism between language and visual signals. In addition, we also proposed the contrastive pre-training step to reduce the domain gap between the image encoder and the target dataset. To test the generalisability of our learning scheme, we collected and verified our model on the world-first multi-modality retina report generation dataset referred to as Retina ImBank and another large-scale retina Chinese-based report dataset referred to as Retina Chinese. These two datasets will be made publicly available and serve as benchmarks to encourage further research exploration in this field. From our experimental results, we demonstrate that our proposed method has outperformed multiple state-of-the-art image captioning and medical report generation methods on IU X-RAY, MIMIC-CXR, Retina ImBank, and Retina Chinese datasets.

Keywords: Medical report generation, Vision and language.

1. Introduction

Writing medical reports is one of the major routine works for radiologists and ophthalmologists. These medical reports describe observations and diagnostic findings based on the knowledge of medical professionals. However, it is challenging to control the reports' qualities due to the experience variations of medical professionals. Therefore, generating medical reports in a unified standard is an essential process for disease diagnosis and treatment. Besides, proposing reliable and accurate medical report generation methods helps reduce labor-intensive workload [1, 2]. To be more specific, the tedious process of examining medical images and typing findings of diseases and lesions into the computer system can be replaced.

Figure 1: Examples of images and corresponding reports from MIMIC [3] and Retina Chinese datasets.

The image captioning task appeared earlier than the medical report generation task, and both of them have built interaction between vision and language. Inspired by the image captioning tasks [4] in deep learning research, many medical report generation tasks [5, 6, 7] have been proposed. Compared with conventional image captioning tasks, the medical report generation task has unique challenges and difficulties. Firstly, unlike the descriptive sentences for natural images, the diagnostic medical reports, which might consist of *Impression* and *Findings* sections shown in Fig. 1, are more diverse. Both sections can include sentences with varying lengths. Secondly, object location variations in natural images result in a notable region of interest, while majorities of diseases and lesions occupy relatively small regions. The challenges mentioned above increase the difficulty in understanding medical images. Thirdly, the abnormal findings in the medical images are quite rare compared to the normal findings [8]. Low-frequent rare diseases introduce the problem of imbalanced samples.

Due to the significance of proposing automatic medical report generation methods and the challenges mentioned above, some approaches focus on medical reports generation for different medical image modalities, including pathology images [9], diffusion-weighted imaging [10], retinal images [7], X-ray images [8, 11, 12, 13, 14, 2]. Those approaches are generally within a skeleton of encoder-decoder that encodes the input image into a context vector and then decodes the context vector to a sentence.

In a previous research [13], it has been discovered that pre-training the image encoder with large-scale data from the same domain can improve the performance compared with loading ImageNet [15] pre-trained models or randomly initializing. However, it is difficult to obtain both large-scale and labeled data in the medical domain. Some methods select the domain-inconsistent ImageNet pre-trained model as their image encoder [2, 7], and other approaches acquire extra classification labels [1, 8] to pre-train the image encoder. Notably, the unlabeled images in the medical domain are not inaccessible and also contain beneficial information. The medical report generation datasets have already contained a large number of images, and the history records in the healthcare system can also provide unlabeled medical images as auxiliary data. Taking advantage of the unlabeled data is a potential solution for image encoder pre-training.

Based on the above analysis, we propose a contrastive pre-training procedure to enhance the image encoder. By applying contrastive pre-training, there is no domain gap between the image encoder and the evaluated dataset. In contrastive pre-training, a pretext task enables conventional image classification training without the actual image labels. The medical images are processed into queries, positive keys, and negative keys. The model is expected to discriminate the positive keys from the mixed keys set according to the queries. The resulting parameters of the pre-trained encoder are saved and loaded into the initialized weight of the visual extractor in the medical report generation framework.

In addition to the contrastive pre-training on medical report generation, there are potential improvements by encouraging vision and language fusion via the advanced attention mechanism. The conventional scaled-dot product attention in the Transformer maps three input vectors: query, key, and value to the weighted sum of values. The weighted coefficient of value is obtained by comparing each query-key pair and implemented as the matrix multiplication of query and key. This conventional design ignored higher-order interactions between the query and key-value pairs. In other words, the representative capability of intermediate feature maps generated in the conventional mechanism can still be improved. Therefore, we proposed the linear interaction attention module to introduce second-order interaction between query and key while it reserves the first-order interaction attention. The complex multi-modal relationship between the hidden language features (query) and visual features (key) can be further and better exploited. Therefore, the relationship between low frequent imaging cues such as abnormal regions and medical terms is strengthened with this feature interacting attention module.

To explore the report generation task in the field of ophthalmology, we collected and processed two ophthalmic image-text datasets. The accurate diagnostics of retina diseases, e.g., macular edema, involve different imaging modalities, including fundus photograph (FP) and optical coherence tomography (OCT). The previous ophthalmic image-text dataset [7] is usually based on one or two modalities. We propose the first large-scale and multi-modality ophthalmic image-text dataset **Retina ImBank**, and image modalities include FP, OCT, fundus fluorescein angiography (FFA), fundus autofluorescence (FAF), Indocyanine Green Chorioangiography (ICG), red-free filtered fundus images. We also collected the **Retina Chinese** with 620,215 images and 10,979 Chinese clinical reports from real clinical cases. In addition, we proposed and implemented a practical pipeline for processing sequential FFA images and medical reports in Chinese.

Our contribution can be summarized as follows:

(1) We proposed the novel weighted query-key interacting linear attention module to increase the capability of expressing a complex multi-modal relationship between the visual feature space and the semantic feature space.

(2) We are the first to introduce contrastive pre-training to the medical report generation. We provide a solution to obtain a domain-consistent

image encoder by exploiting the latent information of the dataset itself, which enables the proposed method to be generalized to datasets with multi-modalities.

(3) We collected and processed **Retina ImBank** and **Retina Chinese**, which will be released to serve as benchmarking datasets to encourage further research on generating reports with retina images.

(4) Evaluated on two retina datasets and two Chest X-Ray datasets, the proposed method achieved state-of-the-art performances in majorities of natural language evaluation metrics. Our ablation study shows that the proposed individual module is effective, and the proposed module can improve the performance of the baseline Transformer significantly, e.g., with contrastive pre-training, the BLEU-1 score improved from 0.396 to 0.462 on IU X-RAY.

2. Related Work

Most medical report generation approaches [1, 9, 10, 7, 8, 12, 13, 14, 2] consist of an image encoder and a language decoder. In this section, we first review the existing methods for the image encoder component. Then we further elaborate on the contrastive learning approaches for self-supervised pre-training. Lastly, we review the language decoders in the medical report generation task.

2.1. Image Encoder

Depending on the format of the context vector generated by the encoder, current encoding approaches can be categorized into two types. Some approaches [12, 16, 14] explicitly classify the images and obtain abnormalities or predicted diseases. This strategy is suitable for the template-retrieval-based approaches. On the other hand, some approaches [2, 8, 7, 17, 18] encode the images by extracting image features, which are suitable input for RNN-based model [8, 7, 17, 18] and Transformer-based models [2]. They usually use the output of the last pooling layer in CNNs as the image features. There are also works [1, 13] integrating both the two-image-encoding mode to obtain a more comprehensive image feature representation.

Pre-training the image encoder is a general strategy to improve performance. Some works [2, 7] choose to load the ImageNet [15] pre-trained model, which is publicly available and widely used in image captioning, VQA, and transfer learning. However, Raghu et al. [19] showed that transferring from ImageNet pre-trained benefits little to performance in the medical image classification tasks. The alternative option is pre-training the image encoder with the target dataset. To perform general CNN training with the cross-entropy loss, some works [7, 1] extract the labels from the target dataset's reports automatically or manually. Another option is using auxiliary datasets. For chest X-ray modality, there are already several classification datasets [20, 21] focusing on extracting labels from reports. Thus some approaches [12, 8, 13, 16] select those datasets to pre-train their image encoders.

However, all previous studies are based on the supervised-learning technique requiring annotated labels to train the image encoder and can not be applied directly to newly collected datasets. Thus, self-supervised learning is a potential solution to provide a domain-consistent image encoder for datasets whose labels are difficult to acquire.

2.2. Contrastive Learning

Contrastive learning is a rapidly growing field in self-supervised image representation learning. The SimCLR [22] is a typical contrastive learning framework in which images are prepared into similar (positive) and dissimilar (negative) pairs, and the model is trained to discriminate the negative pairs against the positive pairs. The Momentum Contrast (MoCo) [23] introduces a momentum mechanism that maintains the negative keys in a queue. Grill et al. [24] proposed a BYOL (Bootstrap Your Own Latent) approach, achieving state-of-the-art performance without using any negative pairs. Chen et al. [25] proposed SimSiam learning image representation without negative sample pairs, large batches, and momentum encoders. Zhou et al. [26] proposed the C2L (comparing to learn) method, which outperforms the previous state-of-the-art approach to the chest X-ray image classification task.

Chen et al. [27] compare MoCo and SimCLR in computing cost. The SimCLR needs a large batch size of 4096 to provide sufficient negative pairs and achieve its best performance. It costs $93.0G$ GPU memory in the estimate. In contrast, the MoCo gets its best performance with a batch size of 256, which requires about $5.0G$ GPU memory. Also, the difference in the encoder updating mechanism makes MoCo less costly in terms of training time. Therefore, since the MoCo framework is resource-efficient, we choose MoCo to pre-train the medical image encoder.

2.3. Language Generation

Based on the methodology, the language decoders can be divided into template retrieval-based [12, 14], RNN-based [1, 13, 10], and Transformer-based [2]. The template-retrieval-based methods require building the template database for different datasets with massive manual work involved. On the contrary, the RNN-based or Transformer-based methods can be directly applied to image-text datasets.

The RNN-based approaches are based on the Show&Tell [28] model. As an extension to Show&Tell [28], Xu et al. [29] first introduced an attention mechanism to image captioning. More recently, a collection of works introduced diverse attention mechanisms, including Adaptive Attention [30], Up-down [31], and Attention on Attention [32], X-Linear [33]. These approaches are with an RNN core (LSTM, GRU, e.t.c.) and predict the next word recursively. In the medical domain, both Jing et al. [1] and Liu et al. [18] propose a hierarchical LSTM including a sentence LSTM generating the topic vector and a word LSTM generating individual words.

The Transformer-based approaches do not rely on sequential input. The Transformer encodes the word positions and feeds them into a multi-attention mechanism module. Based on the Transformer, further modifications include extra memory [34, 2], layer connection [34], attention mechanism [33, 35], e.t.c. In recent research, the Transformer-based approaches have reached state-of-the-art in several image captioning and medical report generation tasks, e.g. mesh-memory transformer [34] on MS-COCO captioning [4] and memory-driven transformer [2] on IU X-RAY [5]. R2Gen [2] tackles the medical report generation with the relational memory module and the memory-driven conditional layer normalization. Considering the great efficiency and performance, we select Transformer as our baseline language decoder.

3. Methods

The medical report generation system generates multiple sentences in a certain order, which describe findings and impressions of medical diseases. Each sentence is represented by a set of tokens.

Our proposed method has two steps, the contrastive pre-training and the medical report generation as illustrated in Fig. 2. The contrastive pre-training trains a CNN model to learn image representation from the images training set $I = \{I_1, I_2, ..., I_r\}$. The pre-trained model parameter θ_{CP} is stored for the following steps.

The report generation follows standard encoder-decoder structure [36]. The input is a single image or multiple images I, and the output is the corresponding report $S = \{y_1, y_2, ..., y_o\}, y \in \mathbb{T}$, where y denotes the single word, o represents the total number of words in the report S, and $\mathbb{T}$ is the token set.

The report generation process can be formatted as $X = f_{\theta_{cp}}(I)$ and $Y = f_{\theta_d}(X)$. Here $X = \{x_1, x_2, ..., x_s\}, x \in \mathbb{R}$ refers to the extracted image features where the x is patch-based image feature. The $f_{\theta_{cp}}$ is the image encoder loading the weights of θ_{CP} from contrastive pre-training and f_{θ_d} is the language decoder. The probability of generating the medical report by combining multiple sentences into complete and single targeting sequences is computed as:

$$p(S|I, \theta_{CP}, \theta_d) = \prod_{o=1}^{O} p(y_o|y_1, y_2, ..., y_{o-1}) \tag{1}$$

The objective of the medical report generation task is to produce the medical report maximizing the negative conditional log-likelihood of S:

$$\theta = \operatorname*{argmax}_{\theta} \sum_{o=1}^{O} \log p(y_o | y_1, y_2, ..., y_{o-1}, I; \theta) \tag{2}$$

where $S = \{y_1, y_2, ..., y_o\}$ represents the targeting sequence and o is the maximum number of tokens of a single report.

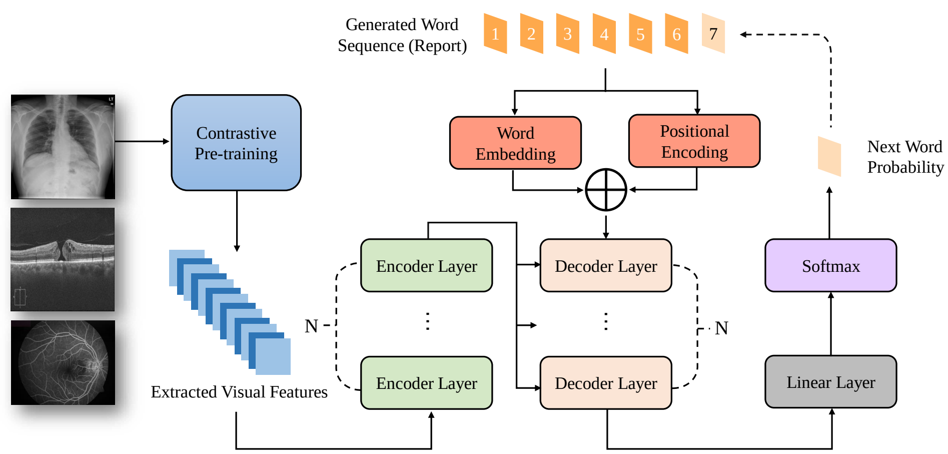


Figure 2: Illustration of the overall architecture. + denotes the add operation, and positional encoding [36] introduces relative location information of the individual feature token in the whole sequence.

3.1. Contrastive Pre-training

As discussed in Section 2.2, the MoCo [23] has advantages in computing and performance. Hence in this paper, we choose the MoCo v2 [27] as our contrastive learning method[2]. The Fig. 3 illustrates the overall framework of MoCo [23].

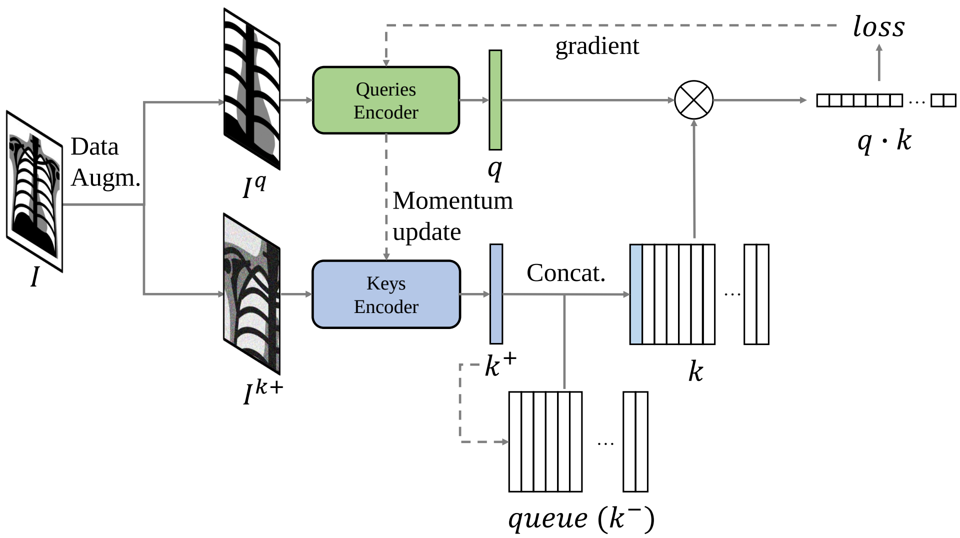


Figure 3: The detailed MoCo (v2) [23] framework. The *Augm.* stands for *augmentation*. The *Concat.* stands for *concatenation*. The solid arrow lines mean the forward operations, and the dashed arrow lines mean the backward operations.

The MoCo treats contrastive learning as a dictionary look-up problem. In the forward direction, the input image I is processed by two random data augmentations and outputs two "views" I^q and I^{k^+}. Then I^q and I^{k^+} are respectively encoded by queries encoder (f_q) and keys encoder (f_k) into a query (q) and a positive key (k^+) and followed by the normalization operation. The encoders mentioned above can be any CNN but have to be in the same architecture. The positive key is concatenated with the negative

[2]MoCo v2 is an upgraded edition of MoCo and keeps the same framework, so we still refer to it as MoCo

keys maintained by a queue containing the previous keys. With the query (q) and the keys (k), the InfoNCE loss [37] can be computed as:

$$L_{q,k^+,\{k^-\}} = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum\limits_{k^-} \exp(q \cdot k^-/\tau))} \tag{3}$$

where q is a query representation, k^+ is positive (similar) key representation, $\{k^-\}$ are a set of negative (dissimilar) key representations and τ is a hyperparameter of temperature.

During the encoder updating procedure, the queries encoder is updated in the conventional back-propagation while the keys encoder is updated by the momentum updating principle as:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{4}$$

where θ_k and θ_q are the parameters of keys encoder and queries encoder, and $m \in [0, 1)$ is a hyperparameter momentum coefficient. At last, the positive key (k^+) is pushed into the queue and replaced with the earliest key in a FIFO (first-in, first-out) manner. After the contrastive pre-training procedure, parameters of the queue encoder θ_q are stored to be the initial weights of the image encoder in the report generation system.

3.2. Transformer Structure

The Transformer structure consists of encoder layers and decoder layers. The encoder layer is composed of Linear Interaction Multi-Head Attention Layer (LIMHA), Add & Norm Layer (ANL), Linear Layer, and another ANL. The purpose of the Linear Interaction Multi-Head Attention Layer is to improve the representative capability of intermediate features by providing second-order or higher-order interactions between the query, key, and value matrices. At the decoder layer, the report is generated in a "shifted right" manner, which uses the known output to predict the next word and is denoted as:

$$y_t = f_{\theta_d}(H, (y_1, y_2, ..., y_{t-1})) \tag{5}$$

where the H is the hidden feature from the immediate layer. The known output is added with the positional encoding, which embeds the positional information of the sequence. It is calculated by sine and cosine function with word position (pos), model dimension (d_{model}) and embedding dimension (i):

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{6}$$

The decoder layer is composed of Masked LIMHA, ANL, Linear Layer, and LIMHA as shown in Fig. 4.

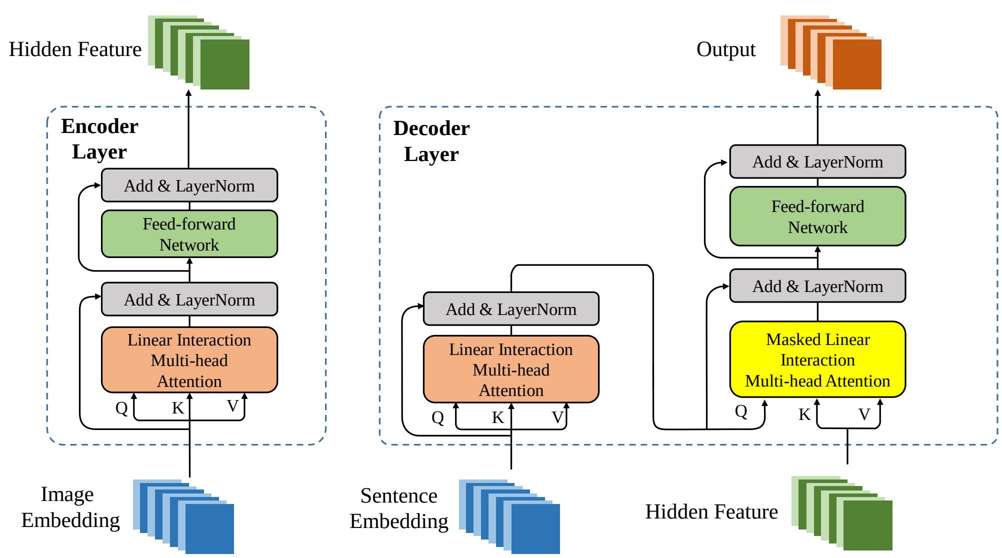


Figure 4: Detailed illustration of the proposed encoder and decoder.

3.3. Linear Interaction Attention Mechanism and Linear Interaction Multi-Head Attention

There are feature dimension differences between medical images and diagnostic reports. It is difficult to associate regions of interest in the medical images with feature maps of corresponding reports. Thus, the weighted query-key interaction Linear Interaction Attention (LIA) mechanism is designed and shown in Fig. 5. The inputs of the attention module include keys (K), values (V), and queries (Q). In the encoder-decoder attention layers, the keys and values are from the output of the encoder, and the queries are from the previous decoder layer. In the self-attention layer, the keys, values, and queries are all from the previous layer. The Linear Interaction Attention mechanism, which describes the mapping relationship between the query matrix and key-value matrices, is defined as follows:

$$\begin{aligned} K^1, V^1, Q^1 &= NN_1(K), NN_2(V), NN_3(Q) \\ K^2, V^2, Q^2 &= NN_4(K^1), NN_5(V^1), NN_6(Q^1) \\ Scores &= f_{mask}(\alpha(K^2 \otimes Q^2) + \beta(K^1 \otimes Q^1)) \\ LIA(K, Q, V) &= f_{softmax}(Scores) \otimes V^2 \end{aligned} \tag{7}$$

where LIA, K, Q, V, NN, and $\otimes$ represent Linear Interaction Attention, key matrix, query matrix, value matrix, linear layer, and element-wise matrix multiplication, respectively; f_{mask} fills 1.0×10^{-9} where the mask template is True; α and β are coefficients to balance the contribution of second-order interacting attention and first-order attention. The design of Linear Interaction Multi-Head Attention is to improve the feature representation capability in the subspace. The computation of Linear Interaction Multi-Head Attention (LIMHA) is defined as:

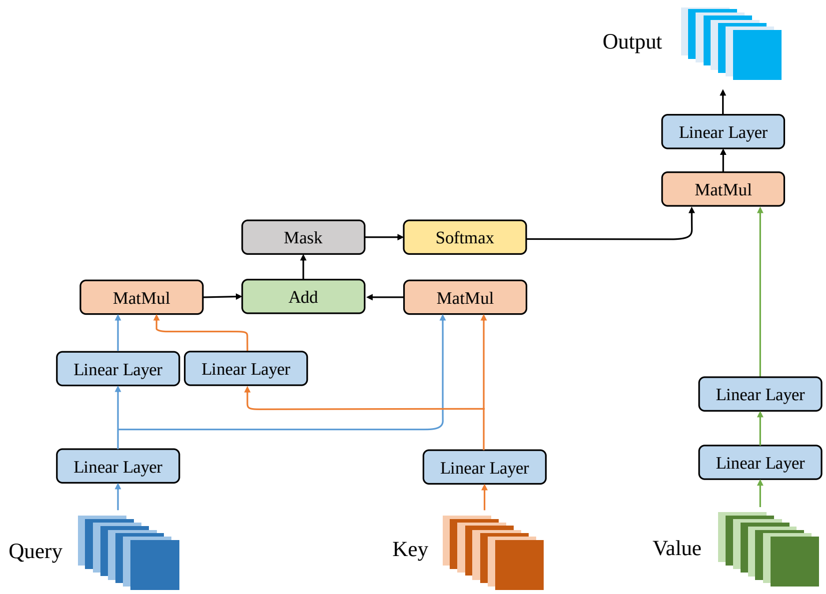


Figure 5: Detailed illustration of linear interaction attention. Q, K, V represent the query matrix, key matrix, and value matrix.

3.4. Beam Searching

Beam searching [38] is also implemented to boost the standardization and quality of generated medical reports. Predicted outputs are sequences rather than simple classification results. The sequence of probabilities computed by multiplying the candidate probability together should be maximized. The Beam searching algorithm defines the beam size, the number of beams for parallel searching. The greedy search algorithm is a special case of the beam searching algorithm, which only selects the best candidate at each step, which might result in a locally optimal choice rather than the optimal global choice. The beam size is bs, and beam searching can be categorized into the following steps. Firstly, the top bs words with the highest probabilities are chosen as bs parallel beams. Secondly, the best bs pairs, including the first and second words, are computed by comparing the conditional probability set. Finally, this process is repeated until a stopping token appears.

4. Experiment and Result

4.1. Datasets Preparation and Description

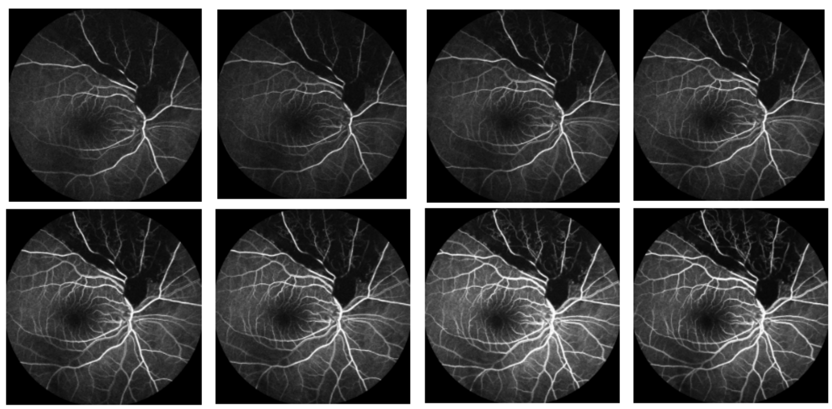


Figure 6: An example of fundus fluorescein angiography image sequences containing similar and repetitive images from the original image dataset without DHash thresholding.

4.1.1. IU X-RAY and MIMIC-CXR

The IU X-RAY [5] and the MIMIC-CXR [3] contain chest X-RAY images and clinical reports. Chest radiography is routinely applied to examine the chest, such as identifying acute and chronic cardiopulmonary medical diseases or conditions. For the IU X-RAY dataset, all the images have frontal and lateral views, and this dataset consists of 6471 images and 3336 reports. MIMIC-CXR is the largest available dataset of chest radiographs with diagnostic reports, which includes 368,960 images and 206,563 reports. For IU X-RAY and MIMIC-CXR dataset experiments, the splits of training, validation, and test follow the R2Gen [2].

4.1.2. Retina ImBank and Retina Chinese

The Retina ImBank dataset has 18,788 retinal images from Retina Image Bank[3] and text captions obtained by us. The modalities of the Retina ImBank dataset include but are not limited to FP, OCT, FFA, FAF, ICG, and red-free filtered fundus images. Each image is associated with one corresponding medical report. The reports basically include information regarding image modalities and the main type of ophthalmic diseases. In addition, some reports include the subtype and detailed findings related to lesions or interesting regions. The reports are verified by three ophthalmologists. The dataset is split into 13146 training images, 1,876 validating images, and 3,755 test images. The maximum and average lengths of reports in the Retina ImBank dataset are 45 and 8.6 tokens, respectively.

The Retina Chinese dataset was collected from real patient cases. Following the MIMIC-CXR dataset, all protected health information of the Retina Chinese dataset was removed before the experiments were conducted. The image modalities of the retina dataset consist of FP, FFA, and ICG. The whole retina dataset is collected by FF450 plus camera (Carl Zeiss Meditec, North America). The original retina dataset without sampling contains many repetitive and similar sequential FFA images shown in Fig. 6. In order to select representative images in the dataset, the hash difference threshold (using 0.6) is performed. After the DHash thresholding strategy, 620215 images were reduced to 57498 images. The word and sentence blocklists are manually created to remove sentences with time descriptions, left/right descriptions,

[3]https://imagebank.asrs.org/

and modality information. Jieba Chinese[4] text segmentation is used to group a few adjacent Chinese characters with medical meanings together. Examples of top frequent words are "retina", "surroundings", "macular area", and "capillaries". The max length in raw Chinese medical reports is 119.

4.2. Evaluation Metrics

In terms of evaluation metrics, the classic natural language metrics, including BLEU [39], METEOR [40], ROUGE-L [41] are selected to assess the performances. Those scores were originally designed for machine translation and machine summarization tasks which are similar to the report generation task.

The BLEU [39] evaluates the position-independent sequential matching and compares the n-grams candidate and the n-grams reference. The BLEU score is computed as

$$\begin{aligned} BLEU &= BP * exp\left(\sum_{q=1}^{n} w_q log(p_q)\right) \\ BP &= e^{min((1-\frac{len(ref)}{len(pred)}),0)} \end{aligned} \tag{8}$$

where q is the number of n-grams, and w_q is the weight of each n-gram class. The brevity penalty (BP) is a multiplicative factor to penalize the length difference between the two sentences. In this paper, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are used to evaluate the predicted report on the corpus level. The BLEU-3 and BLEU-4 are important for our task, because medical terms are often phrases of 3 or 4 words.

[4]https://github.com/fxsjy/jieba

Different from BLEU, METEOR [40] focuses on the computation of sentence-level similarity and evaluates several hypotheses. It is computed as

$$P = \frac{h}{w_t}, R = \frac{h}{w_r}, F_{mean} = \frac{10PR}{R + 9P}, \rho = 0.5(\frac{c}{u_h})^3, \\ METEOR = F_{mean}(1 - \rho) \tag{9}$$

where P and R represent unigram precision and unigram recall; w_t is the number of unigrams in the candidate sentence; w_r is the number of unigrams in the reference sentence; u_h represents mapped unigrams; c is a set of unigrams adjacent in the hypothesis and the ground truth. In this paper, METEOR is used to evaluate the predicted report on the sentence level and semantic similarity.

ROUGE-L [41] measures the longest common sequence between the candidate sentence X and reference sentence Y, defined as:

$$R_{LCS} = \frac{LCS(X,Y)}{m}, P_{LCS} = \frac{LCS(X,Y)}{n}, \\ ROUGE_L = \frac{(1+\gamma^2) * R_{LCS}P_{LCS}}{R_{LCS} + \gamma^2 P_{LCS}} \tag{10}$$

where m and n represent the length of X and Y; LCS represents the process of finding the longest common sequence between candidate and reference; γ is a hyperparameter controlling the relative weight of R_{LCS} and P_{LCS}. In this paper, ROUGE-L is also used to evaluate the report on the sentence level. The difference to METEOR is that ROUGE-L only considers the recall than precision.

4.3. Experiment Setting

The input images of the IU X-RAY are preprocessed to 512×512. For the MIMIC-CXR, Retina ImBank dataset, and Retina Chinese dataset, the image resolutions in preprocessing are 256×256. For the reports in the IU X-RAY and MIMIC-CXR dataset, the max sequence length is 60 and 100, depending on the average report length. For the IU X-RAY dataset, the word with a lower frequency (WLF) of less than three occurrences is marked as *unk* (Unknown). The final vocabulary (FV) for the IU X-RAY dataset has a size of 727 tokens. Similarly, for MIMIC-CXR, WLF and FV are set to 10 and 3471. For Retina ImBank, WLF and FV are set to 3 and 247. For Retina Chinese, WLF and FV are set to 5 and 2001.

For the image encoders in both the contrastive pre-training and feature medical report generation tasks, the ResNet-101 is selected as its visual-feature extractor. In the visual extractor of medical report generation, the final average output size after the pooling operation is adjusted to output image features with 14×14 by removing the fully connected layer.

For the contrastive learning procedure, the random augmentation setting follows the MoCo v2 and includes Random Resized Crop (to 224×224 size), Color Jitter, Random Grayscale, Gaussian Blur, and Random Horizontal Flip. The learning rate schedule is a cosine learning rate schedule. The optimizer is SGD with weight decay 0.0001 and momentum 0.9. The softmax temperature is set to 0.07. In all our contrastive learning studies, we use the training and validate split from the dataset to pre-train the image encoders. In terms of the IU X-ray, the MIMIC-CXR, the Retina ImBank, and the Retina Chinese datasets, the numbers of the image used in pre-training are 4726, 272918, 16899, and 50423, respectively.

The head number of linear interaction multi-head attention and masked bi-linear multi-headed is set to 8. The number of training epochs is 60 for the MIMIC-CXR dataset and 100 for other datasets. The input and output channel numbers of NN_1, NN_2, and NN_3 are 64. We also apply a beam-searching technique [38] with a beam size of 3 to explore the subsequence with the highest probability. The training optimizer is Adam optimizer with 0.00005 weight decay. The initial learning rates are 0.0001 for the proposed encoder-decoder structure and 0.00005 for the visual feature extractor. The α and β are both set to 0.5 for the ablation study.

4.4. Encoder Pre-training Hyperparameter Analysis

This study is conducted on the IU X-ray dataset to investigate the hyperparameters in the image encoder pre-training procedure. The experiment selects the MoCo key discrimination task accuracy as the benchmark. The hyperparameters include batch size, momentum coefficient, queue size, and temperature. The other hyperparameters follow the default setting of MoCo v2 [27].

Table. 1 shows the key-discrimination task accuracy under the hyperparameter grid-searching. We observe that the accuracy increases with decreasing in the batch size. For small batch sizes, the accuracy increases to over 85%. We selected a large batch size (128) and a small one (8) for the rest of the experiments. We also observe that the decreasing momentum coefficient can cause a 100% accuracy for in whole training period (marked as "fail" in Table. 1). According to Eq. (4), the $(1 - m)$ can be thought of as the "learning rate of keys encoder". When this "learning rate" is high (small m), the keys encoder can "catch up" with queries encoder (the keys encoder is updated each batch). Therefore, the over-similar encoders can lead to extremely high top-1 accuracy under a small momentum coefficient. For the

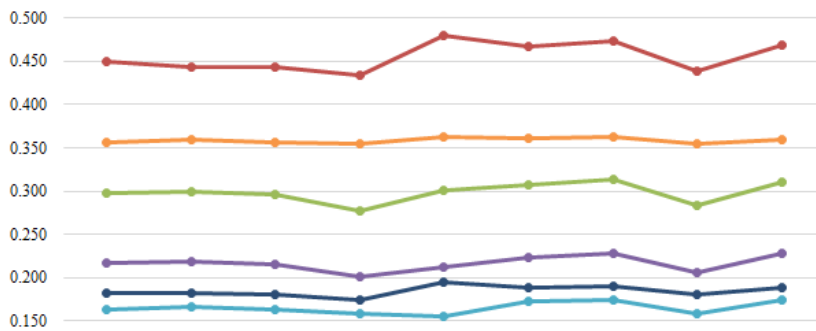
Table 1: MoCo hyperparameter experiments of key discrimination task accuracy. The Acc.1 and Acc.5 are top-1 accuracy and top-5 accuracy of the key discrimination task.

Batch Size	Learning Rate	Momentum	Queue Size	Temperature	Acc.1(%)	Acc.5(%)
128	0.03	0.999	65536	0.07	17.6	34.6
64	0.01	0.999	65536	0.07	51.2	70.2
32	0.005	0.999	65536	0.07	73.9	87.7
16	0.003	0.999	65536	0.07	81.3	92.5
8	0.001	0.999	65536	0.07	86.8	95.4
128	0.03	0.99	65536	0.07	73.8	87.3
128	0.03	0.95	65536	0.07	Fail	
8	0.001	0.99	65536	0.07	88.1	97.3
8	0.001	0.95	65536	0.07	Fail	
128	0.03	0.999	16384	0.07	15.8	32.3
128	0.03	0.999	1024	0.07	38.7	68.3
8	0.001	0.999	16384	0.07	86.8	95.4
8	0.001	0.999	1024	0.07	89.8	98.0
128	0.03	0.999	65536	0.1	12.3	26.4
128	0.03	0.999	65536	0.4	2.9	3.5
128	0.03	0.999	65536	0.7	2.9	3.4

queue size, we find that the different batch sizes have different optimal queue sizes. The queue size determined the number of saved keys in the MoCo framework. A large queue may contain too many keys encoded long ago (up to 14 epochs ago when 65536 for IU X-ray), and a small queue maybe not be diverse enough to represent the whole feature space of the dataset. Also, the batch size will determine the queue updating frequency, affecting the recency of the information in the queue. Therefore, the queue size and batch size have a combined effect on the accuracy. With the study on temperature, we observe the accuracy drops with the increase of temperature τ in the loss function.

4.5. Language Decoder Hyperparameter Analysis

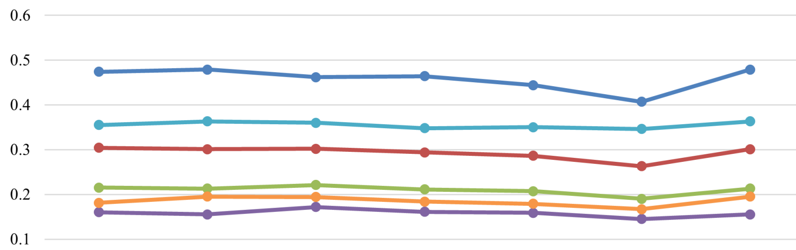
This study is conducted on the IU X-ray dataset to investigate the hyperparameters selection in the language decoder. The experiment uses the language benchmarks, and the hyperparameters include the α in linear interaction attention and the beam size in the beam search technique.



α	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
BLEU-1	0.450	0.443	0.443	0.434	0.479	0.467	0.474	0.439	0.468
BLEU-2	0.298	0.299	0.296	0.277	0.301	0.307	0.313	0.283	0.310
BLEU-3	0.217	0.219	0.215	0.202	0.213	0.224	0.228	0.206	0.228
BLEU-4	0.164	0.166	0.163	0.158	0.155	0.172	0.174	0.158	0.175
ROUGE-L	0.357	0.360	0.357	0.354	0.363	0.361	0.362	0.355	0.360
METEOR	0.183	0.182	0.180	0.174	0.195	0.188	0.190	0.180	0.188

Figure 7: Hyperparameter experiments of first-order and second-order attention. α and β denote the first-order and the second-order attention, respectively.

As defined in Eq. (7), the α and β denote the coefficients of the second-order attention and the first-order attention, respectively. We test the α from 0.1 to 0.9 in an interval of 0.1, with $\beta = 1 - \alpha$, shown in Fig. 7. The $\alpha = 0.5$ shows the best score in BLEU-1, ROUGE-L, and METEOR. Meanwhile, the $\alpha = 0.3$ shows the best score in BLEU-2 and BLEU-3, and the $alpha = 0.1$ shows the best score in BLEU-4. Overall, we can observe that the medium α (0.3-0.5) has better performance than the smaller (0.1-



Beam Size	2	3	4	5	10	20	20
BLEU-1	0.474	0.479	0.462	0.464	0.444	0.407	0.479
BLEU-2	0.304	0.301	0.302	0.294	0.286	0.263	0.301
BLEU-3	0.215	0.213	0.221	0.211	0.207	0.19	0.213
BLEU-4	0.16	0.155	0.172	0.161	0.159	0.145	0.155
ROUGE-L	0.355	0.363	0.36	0.348	0.35	0.346	0.363
METEOR	0.181	0.195	0.194	0.184	0.179	0.167	0.195

Figure 8: Hyperparameter experiments of the beam size.

0.2) or larger (0.6-0.9) in the language metrics. It indicates that second-order attention and traditional first-order attention are both essential to achieve the best performance compared with other methods, and first-order attention has slightly more attribution. The beam size experiment results are shown in Fig. 8. Results indicate that the smaller beam size is suitable for the IU X-ray task. The reason may be the phrases in the IU X-ray are usually short but diverse making the smaller beam size advantageous.

4.6. Baseline Comparison

In order to demonstrate the effectiveness of the proposed linear interaction attention mechanism and contrastive pre-training, ablation studies were performed and shown in Table. 2. By introducing the contrastive pre-training module in all four datasets, all language evaluation metrics increased, respectively, demonstrating the effectiveness of the contrastive pre-training. It indicates contrastive pre-training achieves a better representative ability in

Table 2: Ablation studies and comparison based on different datasets to demonstrate the effectiveness of the proposed components.

Dataset	Image Encoder	Language Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Retina ImBank	RI	Base	0.474	0.365	0.298	0.218	0.523	0.216
	IN	Base	0.497	0.389	0.319	0.237	0.527	0.223
	CP	Base	0.622	0.539	0.482	0.421	0.656	0.318
	IN	Base + LIMHA	0.627	0.544	0.486	0.426	0.666	0.320
	CP	Base + LIMHA	**0.638**	**0.561**	**0.508**	**0.456**	**0.676**	**0.332**
Retina Chinese	RI	Base	0.352	0.225	0.155	0.116	0.296	0.157
	IN	Base	0.354	0.233	0.164	0.126	0.323	0.161
	CP	Base	0.369	0.240	0.168	0.127	0.312	0.163
	IN	Base + LIMHA	0.357	0.244	0.180	**0.143**	**0.338**	0.161
	CP	Base + LIMHA	**0.371**	**0.249**	**0.181**	0.142	0.336	**0.168**
IU X-RAY	RI	Base	0.399	0.258	0.183	0.135	0.352	0.170
	IN	Base	0.396	0.254	0.179	0.135	0.342	0.164
	CP	Base	0.462	0.293	0.201	**0.159**	0.358	0.184
	IN	Base + LIMHA	0.430	0.276	0.198	0.151	0.349	0.176
	CP	Base + LIMHA	**0.479**	**0.301**	**0.213**	0.155	**0.363**	**0.195**
MIMIC-CXR	RI	Base	0.326	0.204	0.138	0.100	0.277	0.130
	IN	Base	0.314	0.192	0.127	0.090	0.265	0.125
	CP	Base	0.348	0.218	0.149	0.109	0.281	0.140
	IN	Base + LIMHA	0.323	0.198	0.132	0.094	0.269	0.126
	CP	Base + LIMHA	**0.362**	**0.227**	**0.155**	**0.113**	**0.283**	**0.142**

RI = Random Initialized ResNet; IN = ImageNet pre-trained ResNet;

CP = Contrastive Pre-training; LIMHA = Linear Interaction Multi-Head Attention.

feature space and is more suitable for medical report generation than pre-training with ImageNet. Experiments are also conducted on every dataset to verify the effectiveness of the linear interaction attention mechanism. The proposed linear interaction attention mechanism improves the language score in all groups.

To compare different contrastive pre-training methods, we conduct a comparison study on the IU X-ray dataset. The compared methods include BOYL [24], SimCLR [22], DenseCL [42], SimSiam [43], MoCo v3 [44], and MoCo v2 [23]. Fig. 9 shows image encoders pre-trained with the MoCo v3 and MoCo v2 can lead to a better language score. 5 of 6 self-supervised ap-

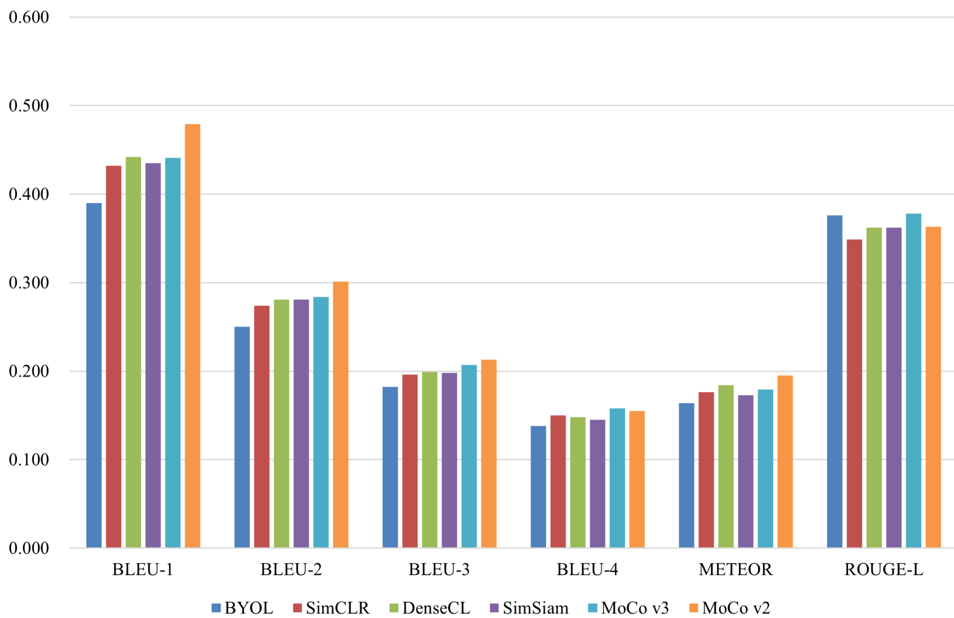
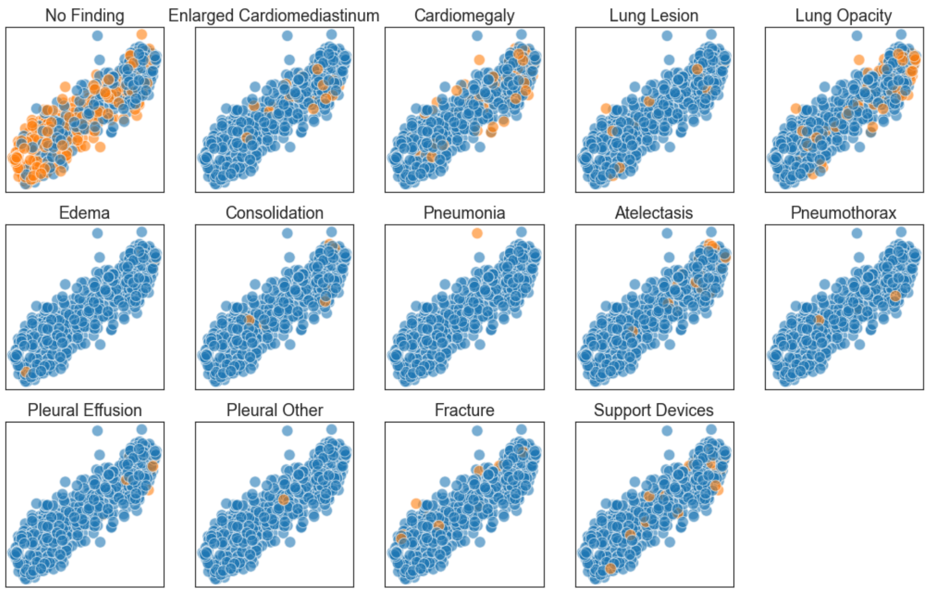


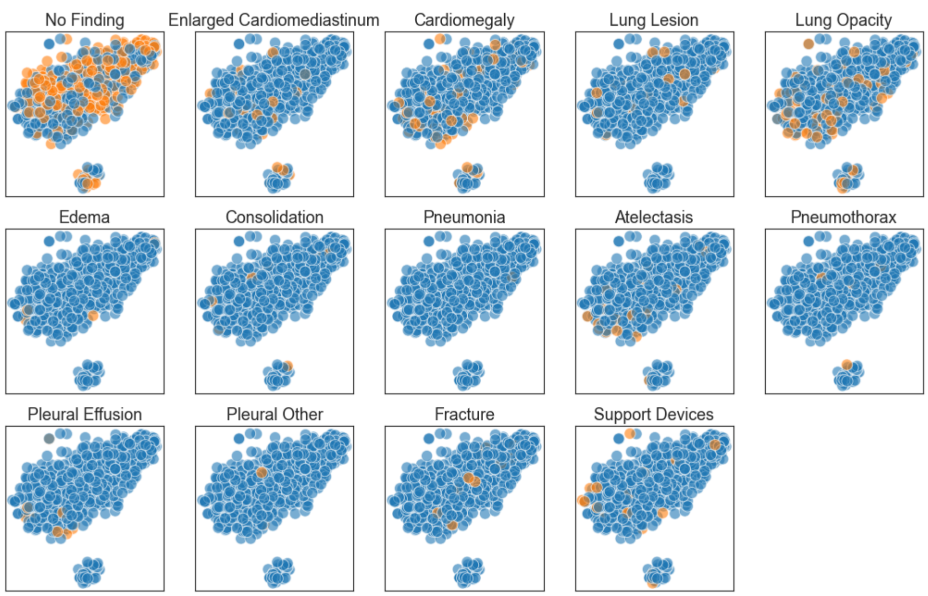
Figure 9: Comparison of report generation language score with different pre-trained image encoder

proaches show better performance than the "Random Initialized" and "ImageNet Pre-trained" in Table. 2. It proves that the domain consistency between the image encoder pre-training procedure and the task domain is an important factor in medical report generation. The contrastive pre-training step can be applied in different contrastive learning schemes.

Especially, we compare a supervised learning encoder on the IU X-ray dataset. We choose the CheXpert labeler [20] to extract 14 labels from the reports and pre-train the image encoder. As shown in Fig. 9, the performance is higher than the ImageNet pre-trained encoder and lower than our contrastive pre-trained encoder. The key problem is that the CheXpert labeler is not accurate enough on IU X-ray reports or not balanced enough to



(a) Supervised Pre-trained



(b) MoCo v2 Pre-trained

Figure 10: The t-SNE visualization on supervised pre-trained and MoCo v2 pre-trained encoders. The orange and blue points represent positive and negative, respectively.

support the pre-training. Moreover, this method is not directly applicable to our Retina ImBank or Retina Chinese dataset because developing a report labeler system for ophthalmology and the Chinese language is much more challenging.

To investigate the encoder difference, we perform the t-SNE visualization of the supervised pre-trained, and the MoCo v2 pre-trained encoders on the IU X-ray test split (the frontal view images only). As shown in Fig. 10, we use the label from the CheXpert labeler to color individual cases represented by discrete points. The MoCo v2 pre-trained encoder significantly separates the points into two groups. However, all 14 labels fail to explain the point clustering, and the clustering may represent other image findings. In comparison, the supervised pre-trained encoder shows almost no separation. The visualization shows that contrastive pre-training is able to improve the initial image feature clustering. Meanwhile, the labels extracted by the CheXpert labeler may not be informative enough to support supervised pre-training.

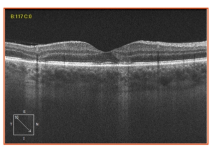
4.7. Case Studies

To further investigate the quality and readability of generated reports, we performed qualitative analysis on three case studies shown in Fig. 11. For the **IU X-RAY Case**, the ground truth report describes three normalities (*pleural effusion, pneumothorax* and *cardiomediastinal silhouette*) and three abnormalities (*low lung volumes, bibasilar atelectasis*, and *thoracic spine*). On the other hand, the medical report generated by the Transformer has correct normal findings and three incorrect abnormalities (*lung volume, bibasilar atelectasis,* and *thoracic spine*). The proposed method is able to provide all three normalities and accurately locates all abnormalities showing the effects

Ground Truth: Low lung volumes with bibasilar subsegmental atelectasis. No focal consolidations pleural effusions or pneumothoraces. Cardiomediastinal silhouette is within normal limits. Degenerative changes of the thoracic spine.

ImageNet Pre-trained: Normal heart size. Bibasilar patchy opacities left greater than right. No pneumothorax or large pleural effusion.

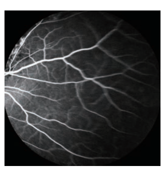
Contrastive Pre-trained: There are low lung volumes. The heart size is mildly enlarged. There are low lung volumes with bronchovascular crowding. There is patchy basilar atelectasis in the left lung base. No large pleural effusion or pneumothorax. Degenerative changes of the thoracic spine.



Ground Truth: Optical coherence tomography showing white dot syndrome astrocytoma specifically acute posterior multifocal placoid pigment epitheliopathy astrocytoma.

Baseline: Optical coherence tomography showing age-related macular degeneration, specifically polypoidal choroidal vasculopathy.

Proposed: Optical coherence tomography showing white dot syndrome specifically acute posterior multifocal placoid pigment epitheliopathy.



Ground Truth: Finding:视盘表面毛细血管扩张渗漏,界限尚清,颞侧、颞下及下方毛细血管扩张,颞下周边部灶性RPE脱色素透见荧光,黄斑区未见明显异常荧光.Impression:考虑葡萄膜炎.

Baseline: Finding:视盘未见明显异常荧光,视盘旁可见一圆形遮蔽荧光区,轻度渗漏,拱环内可见多个点片状低荧光. Impression:CNV.

Proposed: Finding:视盘表面及其上下方和颞上血管弓丛状视网膜新生血管,视盘边界不清,弥漫性视网膜毛细血管扩张渗漏,视盘水肿,视盘以外视网膜血管未见明显异常荧光.Impression:考虑葡萄膜炎.

Figure 11: Illustrations of reports from ground truth, baseline model (Transformer + ImageNet pre-trained CNN), and proposed model for IU X-ray, Retina ImBank, and Retina Chinese. The medical terms are highlighted in different colors.

of better-interpreted imaging features of abnormal regions.

The imaging modality of **the Retina Image Bank Case** is OCT. The abnormalities of this case consist of *white dot syndrome astrocytoma* and *acute posterior multifocal placoid pigment epitheliopathy astrocytoma*. The

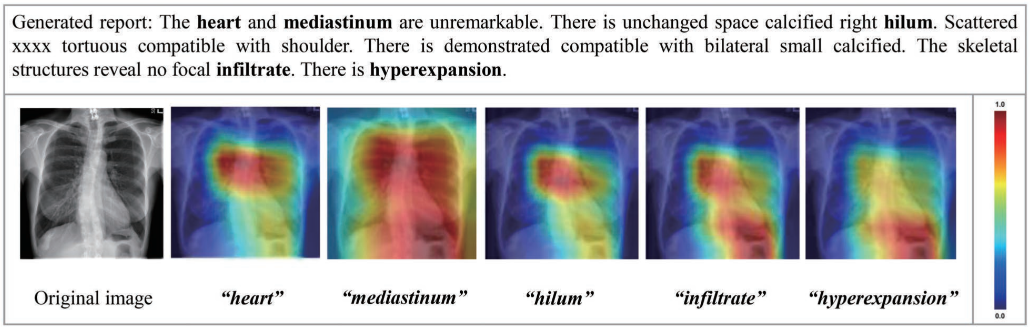
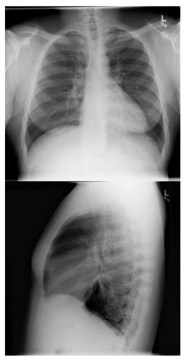


Figure 12: Visualizations of image-text attention mappings of a chest X-ray case from the proposed model. The colors represent the weight strength.

proposed method correctly matchs the retina disease and retina sub-disease. Both the baseline model and the proposed model are able to generate the correct predicted imaging modality. In terms of **Retina Chinese Case**, the *Finding* section describes visual findings of the given retina image, such as the condition of the optic disc, and the impression section relates to the disease diagnostics. The *Impression* section in the medical report of both the ground truth and proposed method is *the possible uveitis*.

To further investigate the model mechanism, we also visualized the attention map of the proposed model. The attention maps are collected from the first cross-attention block where the text is querying the image feature. As shown in Fig. 12, the corresponding image regions of the descriptive words are significantly different and approximately correct despite the limited resolution. It proves that our model has acquired accurate image-text interaction knowledge.



Ground Truth: Cardiomediastinal silhouette is normal in size and contour. Pulmonary vasculature is normal in caliber. Lungs are clear of focal airspace disease pneumothorax or pleural effusion. There are no acute bony findings.

Proposed: The Cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease pneumothorax or pleural effusion. There are no acute bony findings.

Xlinear: The cardiomediastinal silhouette is normal in size and contour. no focal consolidation pneumothorax or large pleural effusion. normal xxxx.

BAN: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax. Osseous structures are within normal limits for the patient age.

Figure 13: Examples of case studies from IU X-RAY of generated medical reports. To further investigate the effects of different attention mechanisms on generated medical reports, qualitative comparison studies are performed. Different colors are chosen to highlight different medical terms.

4.8. Comparison Studies

In this section, we compare the proposed method with state-of-the-art image captioning methods [28, 45, 46, 33] and medical report generation methods [1, 8, 2]. BAN [46] applies bi-linear attention to the object detection feature matrix and hidden language-level information matrix, and X-Linear [33] model attempts to increase feature map representative ability with X-linear attention module implemented by bilinear attention and squeeze-and-excitation. Unlike BAN and X-Linear, the Linear interaction attention mechanism reserves the first-order interaction feature map and removes the over-engineering design of squeeze-excitation. Since the BAN method can not be directly applied to the medical report generation task, we reimplement the bilinear attention mechanism into the proposed Transformer-based framework for comparison. One qualitative example compared with different attention mechanisms is shown in Fig. 13. In this example, X-Linear fails to describe *bony findings*, and BAN can not produce information regarding *pulmonary vasculature*. The proposed method has correct predictions of all normal findings.

Table 3: Comparison study on IU X-RAY dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
ST [28]	0.216	0.124	0.087	0.066	0.306	-
ATT2IN [45]	0.224	0.129	0.089	0.068	0.308	-
ADAATT [30]	0.220	0.127	0.089	0.068	0.308	-
COATT [1]	0.455	0.288	0.205	0.068	0.369	-
HRGR [8]	0.438	0.298	0.208	0.151	0.322	-
BAN [46]	0.453	0.292	0.210	0.096	0.364	0.178
XLinear [33]	0.431	0.270	0.190	0.143	0.344	0.175
R2Gen [2]	0.453	0.288	0.211	**0.165**	0.361	0.182
Proposed	**0.479**	**0.301**	**0.213**	0.155	**0.363**	**0.195**

Table 4: Comparisons study on MIMIC-CXR dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
ST [28]	0.299	0.184	0.121	0.084	0.263	0.124
ATT2IN [45]	0.325	0.203	0.136	0.096	0.276	0.134
ADAATT [30]	0.299	0.185	0.124	0.088	0.266	0.118
TOPDOWN [31]	0.317	0.195	0.130	0.092	0.267	0.128
XLinear [33]	0.332	0.203	0.135	0.096	0.272	0.133
R2Gen [2]	0.353	0.218	0.145	0.103	0.277	**0.142**
Proposed	**0.362**	**0.227**	**0.155**	**0.113**	**0.283**	**0.142**

Table. 3 demonstrates that the proposed method achieved the best performance in all language evaluation metrics on the IU X-RAY dataset except

Table 5: Comparison study on the Retina Chinese dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
ATT2IN [45]	0.269	0.154	0.082	0.053	0.265	0.129
AOA [32]	0.276	0.174	0.117	0.085	0.281	0.139
M2 Transformer [34]	0.298	0.155	0.086	0.058	0.255	0.130
R2Gen [2]	0.309	0.166	0.094	0.063	0.252	0.140
Proposed	**0.371**	**0.249**	**0.181**	**0.142**	**0.336**	**0.168**

BLEU-4. Table. 4 shows the proposed method achieved state-of-the-art results on BLEU scores, and both the ROUGE-L score and METEOR score rank second among seven methods on the MIMIC-CXR dataset.

Since the Retina Chinese dataset is collected from real clinical reports and not released to the public, no comparing methods were evaluated. We selected strong baselines with codes available, including medical report generation method [2] and image captioning methods [45, 32, 34] to compare with. Table. 5 demonstrates the state-of-the-art performance of the proposed method on the Retina Chinese dataset. The medical report generation methods (R2Gen and the proposed) outperform image captioning methods because they are designed for generating sentences with varying lengths. The significant performance improvement of the proposed method is because the image feature encoder obtained in contrastive learning is suitable for describing retina image properties such as abnormal regions and texture.

5. Conclusion

There is no publicly available multi-modality dataset for retina image report generation, so we collected the world-first multi-modality dataset and another retina Chinese dataset to inspire further research on retina report generation tasks. Experimental results also demonstrated that the proposed method generated robust and meaningful medical imaging reports. The linear interaction attention module and contrastive pre-training module improved intermediate feature map fusion capabilities of diseases (or abnormal findings) with imaging cues shown in qualitative and quantitative studies. The contrastive pre-training module was generalized to various datasets without annotated labels. By evaluating with the collected datasets and public chest X-Ray datasets, the proposed method outperformed all comparing methods in majorities of language-matching metrics.

References

[1] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2577–2586.

[2] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1439–1449.

[3] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data 6 (2019) 1–8.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context,

in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[5] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310.

[6] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, arXiv preprint arXiv:1901.07042 (2019).

[7] J.-H. Huang, C.-H. H. Yang, F. Liu, M. Tian, Y.-C. Liu, T.-W. Wu, I. Lin, K. Wang, H. Morikawa, H. Chang, et al., DeepOpht: Medical report generation for retinal images via deep models and visual explanation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2442–2452.

[8] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, Advances in neural information processing systems 31 (2018).

[9] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, MDNet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6428–6436.

[10] A. Gasimova, G. Seegoolam, L. Chen, P. Bentley, D. Rueckert, Spatial semantic-preserving latent space learning for accelerated DWI diagnostic report generation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 333–342.

[11] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[12] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6666–6673.

[13] J. Yuan, H. Liao, R. Luo, J. Luo, 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, Springer, 2019, pp. 721–729.

[14] T. Syeda-Mahmood, K. C. Wong, Y. Gur, J. T. Wu, A. Jadhav, S. Kashyap, A. Karargyris, A. Pillai, A. Sharma, A. B. Syed, et al., Chest X-Ray report generation through fine-grained label learning, in: MICCAI, Springer, 2020, pp. 561–571.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma,

Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252.

[16] B. Jing, Z. Wang, E. Xing, Show, describe and conclude: On exploiting the structure information of chest X-ray reports, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6570–6580.

[17] P. Harzig, Y. Chen, F. Chen, R. Lienhart, Addressing data bias problems for chest x-ray image report generation, in: Proc. British Machine Vision Conference, BMVA Press, Durham,UK, 2019, p. 144.

[18] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest X-Ray report generation, in: Proc. Machine Learning for Healthcare Conference, volume 106 of *Proceedings of Machine Learning Research*, PMLR, Ann Arbor, Michigan, 2019, pp. 249–269.

[19] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: Understanding transfer learning for medical imaging, Advances in neural information processing systems 32 (2019).

[20] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 590–597.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[22] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proc. ICML, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607.

[23] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[24] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent: A new approach to self-supervised learning, 2020. `arXiv:2006.07733`.

[25] X. Chen, K. He, Exploring simple siamese representation learning, 2020. `arXiv:2011.10566`.

[26] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, Y. Zheng, Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations, in: International Conference on Medical Image

Computing and Computer-Assisted Intervention, Springer, 2020, pp. 398–407.

[27] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020. arXiv:2003.04297.

[28] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proc. ICML, PMLR, 2015, pp. 2048–2057.

[30] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[31] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[32] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4634–4643.

[33] Y. Pan, T. Yao, Y. Li, T. Mei, X-Linear attention networks for image

captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10971–10980.

[34] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10578–10587.

[35] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. arXiv:2004.05150.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[37] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018). URL: http://arxiv.org/abs/1807.03748. arXiv:1807.03748.

[38] R. Luo, B. Price, S. Cohen, G. Shakhnarovich, Discriminability objective for training descriptive captions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6964–6974.

[39] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[40] M. Denkowski, A. Lavie, Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011, pp. 85–91.

[41] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[42] X. Wang, R. Zhang, C. Shen, T. Kong, L. Li, Dense contrastive learning for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3024–3033.

[43] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.

[44] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9640–9649.

[45] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: CVPR, 2017, pp. 7008–7024.

[46] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, Advances in neural information processing systems 31 (2018).