Reproducing and Extending Multimodal Medical BERT for Visual Question Answering

João Daniel Silva[1]

Instituto Superior Técnico and INESC-ID, Lisboa, Portugal
joao.daniel.silva@tecnico.ulisboa.pt

Abstract. Models for Visual Question Answering (VQA) on medical images should answer diagnostically relevant natural language questions with basis on visual contents. A recent study in the area proposed MMBERT [16], a multi-modal encoder model that combines a ResNet backbone to represent images at multiple resolutions, together with a Transformer encoder. By pre-training the model over the Radiology Objects in COntext (ROCO) dataset of images+captions, the authors achieved state-of-the-art performance on the VQA-Med dataset of questions over radiology images, used in ImageCLEF 2019. Taking the source code provided by the authors, we first attempted to reproduce the results for MMBERT, afterwards extended the model in several directions: (a) using a stronger image encoder based on EfficientNetV2, (b) using a multi-modal encoder based on the RealFormer architecture, (c) extending the pre-training task with a contrastive objective, and (d) using a novel loss function for fine-tuning the model to the VQA task, that specifically considers class imbalance. Exactly reproducing the results published for MMBERT was met with some difficulties, and the default hyper-parameters given in the original source code resulted in a lower performance. Our experiments showed that aspects such as the size of the training batches can significantly affect the performance. Moreover, starting from baseline results corresponding to our reproduction of MMBERT, we also show that the proposed extensions can lead to improvements. The source code associated to our tests is given at `https://github.com/DannielSilva/MMBERT`.

Keywords: Medical Visual Question Answering · Transformer Encoders · Vision and Language · Computer Vision · Natural Language Processing

1 Introduction

Visual Question Answering (VQA) is an exciting research problem that combines natural language processing and computer vision techniques, currently attracting a significant interest. Some previous efforts have looked into VQA in the context of radiology images [2, 3], where systems can provide additional insights to clinicians, or help patients in the interpretation of their medical images.

A particular formulation for the VQA problem involves taking a medical image (e.g., a radiology image) accompanied by a clinically relevant question, and

producing as output a relevant answer based on the image contents. State-of-the-art approaches for addressing the VQA problem are based on deep neural networks, with systems often following an architecture that features: a first encoder component corresponding to a pre-trained convolutional neural network that generates a representation for the image; a second encoder corresponding to a recurrent neural network or, more recently, a Transformer, that is used to generate a representation for the question; a fusion and classification component that combines both types of information and selects an answer from a set of candidates, thus treating the VQA problem as a multi-label classification task.

Considering the aforementioned formulation, this work reproduces and extends the Multimodal BERT (MMBERT) study [16], which proposed a multi-modal encoder that achieved state-of-the-art performance on the VQA-MED dataset from ImageCLEF 2019 [3]. In this architecture, the visual representations are obtained with a ResNet backbone at multiple resolutions, which are then combined with text representations through a Transformer encoder. The authors leverage the Radiology Objects in COntext (ROCO) dataset, containing medical images paired to textual captions, for model pre-training with a Masked Language Modeling (MLM) objective.

Using the original source code provided by the authors, we first attempted to reproduce the results for MMBERT. This task was met with some difficulties, namely due to the fact that hyper-parameters such as the batch size, the learning rate, or the stopping criteria, can have a strong influence on the results. After this, we extended the model in several directions:

- Using a stronger image encoder, namely the recent EfficientNetV2 [32];
- Using a multi-modal encoder based on the RealFormer architecture [13];
- Extending the pre-training task with a contrastive learning objective [17];
- Addressing the class imbalancement with a recently proposed asymmetric loss function that targets this problem [4];

The rest of this document is organized as follows: Section 2 introduces work related to VQA in the medical domain, while Section 3 describes the MMBERT approach [16]. Section 4 details the different extensions proposed in this work. Section 5 presents the experimental results. Finally, Section 6 presents the main conclusions and proposes directions for future work.

2 Related Work

State-of-the-art approaches for VQA are based on deep neural networks [30]. In most methods, an image encoder corresponding to a pre-trained convolutional neural network generates a representation for the image, while a text encoder corresponding to a recurrent neural network, or more recently a Transformer, generates a representation for the question. The image and question representations can then be fused, informing the generation of the answers. Although answer text can be produced using a decoder component that generates the answer word-by-word [25, 9], most previous studies treat VQA as a classification problem, where each possible answer is seen as a distinct label.

In the medical domain, the VQA task can involve radiology images and clinically relevant questions, e.g. related to the presence of abnormalities or clarifying the organs being depicted. The VQA-MED task of ImageCLEF 2019 [3] provided a dataset to test systems in medical VQA, where all questions can be answered from image contents without requiring additional medical knowledge or domain-specific inference. This dataset contains questions of four different clinical categories (i.e., modality, plane, organ system, and abnormality), constituting an important benchmark in the area.

The best result in the VQA-MED task of ImageCLEF 2019 (i.e., an accuracy of 62.4% and BLEU score of 64.4%) was achieved by a system that combined a VGG16 network to extract image features, with BERT [10] for representing the question features [34]. A co-attention mechanism [35] was used to combine the image and text features, which were fed to feed-forward layers that produce the answer, through a classification objective. Competitive results were achieved with the method reported by Mihn et al. [14], which used an ensemble approach with variations on multiple VQA components. The authors used a ResNet-152 model, pretrained on the ImageNet dataset, to extract visual features, together skip-thought vectors [18] and different BERT [10] implementations to represent the question. The representations of both modalities were fused with an attention mechanism to retrieve global image features, based on Multi-modal Compact Bilinear (MCB) pooling [11], which uses a Tucker matrix decomposition [5] to reduce the dimensionality of the fused feature vector. The question features are further processed through a non-linear transformation to produce global question features. A bilinear transformation is performed on these global features, and the result is finally fed to a feed-forward component to produce an answer.

The use of multi-modal Transformers [33] and self-attention mechanisms has also been explored to unify vision and language [6], through methods like UNITER [8] or ViLBERT [19]. These methods can either follow a single-stream approach, receiving a concatenation of the visual and text tokens that are jointly processed by a single encoder that uses attention over both modalities, or a dual-stream approach, in which the inputs of each modality are encoded separately by Transformer layers, before being jointly modeled in cross-modal Transformer layers. Similar approaches have been used in the medical VQA domain [16, 25], particularly in the VQA-MED task of ImageCLEF 2019 [3]. For instance, Ren and Zhou [25] proposed a method named CGMVQA, where a ResNet152 model is used to extract five visual tokens at 5 different depths. The visual and text tokens are then concatenated and fed to a shallow Transformer. A feed-forward layer is used over a representation obtained from the text tokens to retrieve the type of question. Depending on the type of question, the answer is obtained either with a classification or a language generation decoder. In the classification case, a final classification layer on top of the representation for the first token is used to obtain the answer. The language generation decoder was used for answers of the abnormality category, and the model auto-regressively predicts the next token until the end of the sentence. Based on the previous architecture, the method described in [16], which the authors named MMBERT, addresses the

VQA problem in a classification setting, and also leverages the pre-training of a Transformer with multi-modal medical data. Specifically, the authors propose the use of a masked language modeling objective, in which the model predicts the masked tokens based on the remaining text and image features. MMBERT achieved state-of-the-art results on the VQA-MED dataset from ImageCLEF 2019, and the method is described in more detail next.

3 The Multimodal Medical BERT (MMBERT) Model

The Multimodal Medical BERT (MMBERT) [16] is a multi-modal Transformer architecture designed to address medical VQA. A ResNet152 encoder, pre-trained on ImageNet, is used to capture image representations at 5 different resolutions, while a shallow Transformer is used to model the interaction between visual features and the question. Convolution and global average pooling operations are performed over the results from five different blocks of the ResNet architecture, to obtain 5 visual tokens with the same dimensionality of the Transformer. To obtain text representations, BERT [10] embeddings are first used. The text is tokenized into WordPieces and each token is represented with the sum of token, position, and segment embeddings from a pre-trained BERT model. The tokens of each modality are concatenated into a representation $\{\text{CLS}, \text{img}_1, ..., \text{img}_5, \text{SEP}, c_1, ..., c_N, \text{SEP}\}$, where $\text{img}_1, ..., \text{img}_5$ corresponds to the visual features and $c_1, ..., c_N$ represents the text features. The Transformer encoder model has 4 blocks, a hidden state size of 768, and 12 attention heads, using the ReLU activation function [21] after the feed-forward layers.

The authors used the Radiology Objects in COntext (ROCO) dataset [23] of medical images and their corresponding captions, for model pre-training. This dataset contains over 81k images of several medical imaging modalities. Each image is accompanied by its caption and, additionally, with keywords extracted from the caption, the corresponding UMLS semantic types, and UMLS concept identifiers. The model is pre-trained with a Masked Language Modeling (MLM) objective, which consists on predicting the original tokens of masked positions based on the remaining text and image features. To improve the performance on medical data, only medical keywords are masked and common words are left untouched, leveraging the keyword information of the dataset.

For fine-tuning to the VQA-MED task, instead of taking the standard approach of using the `[CLS]` token representation from the last layer of the Transformer, the representations of each token in the last layer are averaged and fed to a feed-forward component to obtain an answer classification. The authors report an accuracy of 62.4% and a BLEU score of 64.2% with the aforementioned approch, at the same time also reporting that the results can be improved by having 5 independent models for each type of question category, in this case achieving 67.2% in terms of overall accuracy and 69.0% for BLEU.

4 The Proposed Extensions over MMBERT

Besides replication, this study also tried to advance over MMBERT in several directions. For instance, regarding the image encoder, we experimented with the

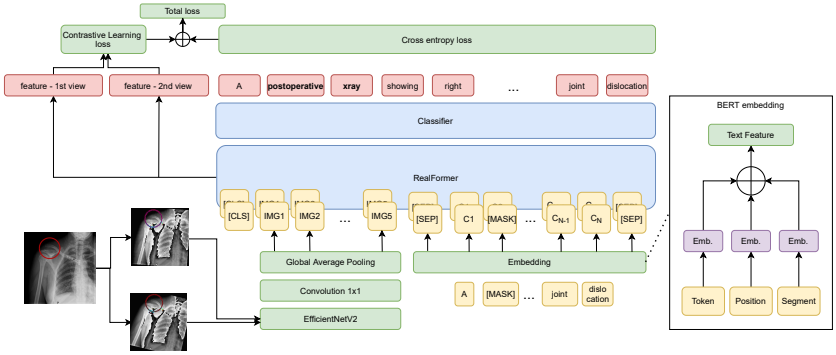

Fig. 1. Graphical depiction for the MMBERT architecture, together with our extensions and featuring the MLM and SimCLR-based contrastive pre-training tasks.

use of EfficientNetV2 [32], which is a recent CNN architecture that achieved state-of-the-art performance in several image classification tasks. In the multi-modal encoder part of the model, we tested the use of a RealFormer [13], which is an extension of the Transformer architecture [33] that creates a residual path between the attention scores of each layer. The ReLU activation function was also replaced by SERF [20], which addresses some drawbacks of the ReLU and achieved promising results on various types of tasks with different models.

Following the MMBERT approach, the visual features are captured at five different depths of the EfficientNetV2, resulting in five visual tokens representing different resolutions. As for the text features, we leverage BERT [10] to first tokenize the text into WordPieces and then to create embeddings. The model is pre-trained with the same masked language modeling objective, although combined with a new pre-training task that leverages contrastive learning. With this method, the model is given a batch of image+caption pairs where each pair has 2 views. Each view is the result of data augmentation from the same original data: image transformations are used to produce 2 views of the image, and back-translation is used to augment the caption text. The representations to be contrasted are extracted at the end of the multi-modal RealFormer, by averaging the representations for each token and passing them to a feed-forward layer with the SERF activation function, to produce features of size 128. The training objective involves pulling together the multi-modal representations from the same original sample, and pushing apart the views from different samples. We experimented with the use of the loss function from SimCLR [7], and also with a supervised contrastive loss [17]. Given the loss for the contrastive task $\mathcal{L}^{con}$ and the cross-entropy loss for the masked language modeling $\mathcal{L}^{MLM}$, the final pre-training loss $\mathcal{L}^{tot}$ is a simple addition of the two:

$$\mathcal{L}^{tot} = \mathcal{L}^{con} + \mathcal{L}^{MLM}. \tag{1}$$

For the downstream task of VQA, we also experimented with an asymmetric loss [4] which addresses the class imbalance present in our classification task. Figure 1 shows a graphical depiction for the pre-training process.

4.1 The EfficientNetV2 Image Encoder

EfficientNetV2 [32] is a family of convolutional neural networks that improved upon EfficientNet [31] with faster training speed and better parameter efficiency. Neural architecture search was used to build these models, in which the key components are the MBConv [26] and Fused-MBConv [12] blocks.

MBConv blocks consist on applying a 1×1 convolution to increase the number of channels in an input representation and then a depthwise 3×3 convolution in the high dimensional space. Then, a 1×1 convolution is used to reduce the number of channels to the original size, so that this final feature map can be added to the input in a residual connection. Depthwise convolutions have fewer parameters and FLOPs than regular convolutions, but cannot be processed on modern accelerators in their full capacity. The authors proposed the Fused-MBConv block to address this limitation, replacing the depthwise 3×3 convolution and expansion 1×1 convolution, in MBConv, with a single regular 3×3 convolution. Neural architecture search is used to automatically search for the best combination of these two building blocks, together with squeeze-and-excitation operations [15]. The authors also proposed to train the model with a progressive learning technique, which consists on gradually increasing the image size as the training progresses, while also increasing the strength of regularization techniques such as dropout [29] and image augmentations.

The EfficientNetV2 used in our experiment has a medium size, with $\approx 54M$ parameters, which is similar to the number of parameters of ResNet152. The model was pre-trained on ImageNet-1k with progressive learning. The model and pre-trained weights were provided in the `timm` Python package[1].

4.2 The RealFormer Multi-Modal Encoder

The original Transformer architecture [33] follows an encoder-decoder structure based on self-attention. The encoder is composed by a stack of layers (originally $N = 6$) and each layer is composed by a stack of sub-layers, consisting of a multi-head attention module and a feed-forward module. A residual connection is also applied around each sub-layer, followed by normalization.

The attention is computed with a scaled dot-product attention operation, which corresponds to a weighted sum of values in V (i.e. token representations), where the weights are computed by a function of queries Q and keys K:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{2}$$

In the previous equation, $\frac{QK^T}{\sqrt{d_k}}$ represents the raw attention scores for each (query,key) pair, while d_k represents the dimensionality of queries and keys.

[1] `https://github.com/rwightman/pytorch-image-models\#may-14-2021`

The attention module performs an aggregation of the attention calculation over multiple heads. The resulting computation of all heads is concatenated and linearly transformed by a matrix W^O. This process can be formally defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_i, ..., \text{head}_h)W^O, \tag{3}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, Q and K are matrices with dimensionality d_k, and V is a matrix with dimensionality d_v. The parameters W_i^Q, W_i^K, and W_i^V are matrices that linearly project queries, keys, and values into the attention space of the i-th head.

In the feed-forward module, non-linear transformations are performed, using activation functions like ReLU [21]. Layer normalization modules are also used above each sub-layer, to stabilize training.

The RealFormer [13] follows the same general design, adding skip edges to connect multi-head attention modules in adjacent layers. The pre-softmax attention scores from a previous layer are additional inputs to the current one:

$$\text{ResidualMultiHead}(Q, K, V, \text{Prev}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O\,. \tag{4}$$

The variables head_i correspond to $\text{ResidualAttention}(QW_i^Q, KW_i^K, VW_i^V, \text{Prev}_i)$ operations, and Prev_i is a slice of the previous layer corresponding to head_i. A residual attention operation adds the residual score on top of $Prev_i$ and then the regular weighted sum is computed as usual:

$$\text{ResidualAttention}(Q, K, V, \text{Prev}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \text{Prev}\right)V\,. \tag{5}$$

Finally, the pre-softmax attention scores of the current layer, corresponding to $\frac{QK^T}{\sqrt{d_k}} + \text{Prev}$, are passed over to the next layer.

Regarding the parameters of the architecture used in our tests, the model has 4 encoder layers, a hidden state size of 768, and 8 heads for the multi-head attention module. The original ReLU activation function was replaced by SERF (Softplus ERror activation Function) [20], which is a recently proposed activation function that is smooth, non-monotonic, and fully differentiable, avoiding issues with the gradient-based optimization process. Given the Gauss error function $\text{erf}()$, SERF can be formally defined as:

$$\text{f}(x) = x \times \text{erf}(\ln(1 + e^x))\,. \tag{6}$$

Because of numerical stabilization issues in the calculation of the log and exponential functions, the value of x is clipped if it surpasses a threshold before the calculation of the exponential, to prevent overflow. It is also noted that for:

$$c \gg 1, \text{f}(c) = c \times \text{erf}(\ln(1 + e^c)) \approx c \times \text{erf}(c) \approx c \times 1 = c\,.$$

4.3 Contrastive Pre-Training

Contrastive learning is an attractive method for initializing models operating with medical data, due to the much smaller size of the available annotated datasets, when compared to other domains. In this work, we experimented with two contrastive strategies for model pre-training, namely SimCLR and SupCon.

In both cases, the transformations used for image augmentation include cropping, affine transformations, and color jittering. Back-translation was applied to augment the text, where the captions are first translated to Spanish, German and French, and then translated back to English. Transformer models based on Marian-NMT were used in pre-processing to obtain the translated text[2]. Each batch consists of N pairs of image+caption, where each pair has 2 views. For informing the supervised contrastive loss, a similarity matrix can be computed with basis on the captions. Different similarity functions were tested, namely the Jaccard similarity between word tokens, or a semantic metric based on the use of sentence-transformers [24], specifically the model named `all-mpnet-base-v2`[3].

In brief, SimCLR [7], is a simple method for self-supervising learning maximizing the agreement between different transformations of the same input, while minimizing the agreement between transformations of different inputs.

The representations resulting from the last RealFormer layer, before the classifier, are processed with average pooling and fed into a feed-forward layer using the SERF that computes a non-linear projection, producing features of size 128. Considering a mini-batch of N examples, the data augmentations produce two transformations of each instance resulting in $2N$ data points. Considering a transformation pair from the same original inputs as a positive pair, the remaining $2(N-1)$ pairs are considered negative samples. Given a similariy function between two vectors (i.e, the cosine similarity $\mathrm{sim}(u,v) = u^T v/(\|u\|\|v\|)$), the loss function for a positive pair (i,j) is defined as follows:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp(\mathrm{sim}(z_i, z_j)/\tau)} \,. \tag{7}$$

In the previous equation $1_{[k\neq i]} \in \{0,1\}$ is an indicator function that evaluates to 1 iff $k \neq i$. The parameter τ is a temperature parameter. The authors term this loss as the normalized temperature-scaled cross entropy loss (NT-Xent).

Supervised Contrastive (SupCon) learning extends methods like SimCLR to leverage available label information [17], in the form of classes associated to instances or similarities between instances. In SimCLR, only one positive is extracted to pull data augmentations of the same sample together. In SupCon, normalized representations for instances from the same class (or instances known to be similar) are pulled closer together than representations from different classes. The label information allows the use of many more positives per anchor, in addition to the many negatives. The training method is similar to that of SimCLR, where data augmentation is used to obtain two views of the batch, which are

[2] `https://huggingface.co/transformers/model_doc/marian.html`
[3] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

then fed to the model to obtain representations that are re-projected with a feed-forward layer. The loss function is designed to include the information or similarity between instances, generalizing for an arbitrary number of positives. The supervised contrastive loss function is defined as:

$$\mathcal{L}^{sup} = \sum_{i \in I} \frac{-1}{2N} \sum_{p \in I} \log \frac{\text{label}(p,i) \cdot \exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{k \in I} 1_{[k \neq i]} \cdot \text{label}(k,i) \cdot \exp(\text{sim}(z_i, z_k)/\tau)} . \quad (8)$$

Here, $i \in I \equiv \{1, ..., 2N\}$ is the index of an arbitrary augmented sample, $\text{label}(k,i) \in [0,1]$ encodes the similarity between samples k and i (e.g., derived from the textual captions), and $1_{[k \neq i]} \in \{0,1\}$ is an indicator function that evaluates to 1 iff $k \neq i$. The parameter τ is again a temperature parameter, set to 0.07 in our experiments.

SupCon allows every positive in the batch (from augmentation or the other similar instances within the batch) to contribute to the numerator in Equation 8, resulting in a more robust clustering of the representation space. It also preserves the summation over negatives in the contrastive denominator, wherein the ability to discriminate between signal and noise (negatives) can be improved by adding more negative examples.

4.4 Asymmetric Loss for VQA Fine-Tuning

A recent study on image classification proposed an asymmetric loss function [4] for imbalanced multi-label or multi-class classification problems. For multi-class problems, like ours, the main idea is to decouple the processing of the positive and negative cases, assigning them different exponential decay factors.

$$\begin{cases} \mathcal{L}_+ = (1-p)^{\gamma_+} \log(p) \\ \mathcal{L}_- = p^{\gamma_-} \log(1-p) \end{cases} . \quad (9)$$

In the previous equation, $\mathcal{L}_+$ and $\mathcal{L}_-$ are the positive and negative parts of the loss, respectively, p is the networks's output probability vector, and γ_+ and γ_- are the positive and negative focusing parameters, respectively.

For $\gamma_- > 0$, the contribution of the negative part is down-weighted when the probability is low ($p \ll 0.5$). The contribution of the positive part can be emphasized by setting $\gamma_- > \gamma_+$, to help the network learn more meaningful representations. In our tests, we used $\gamma_- = 4$ and $\gamma_+ = 0$, which are the values used by the authors for single-label classification, and given by default in the original source code[4]. The final loss is obtained by combining $\mathcal{L}_+$ and $\mathcal{L}_-$:

$$\mathcal{L} = -y\mathcal{L}_+ - (1-y)\mathcal{L}_- . \quad (10)$$

Label smoothing is also performed, replacing the multi-class hard 0/1 classification targets, with targets of $\frac{\epsilon}{k-1}$ and $1-\epsilon$, respectively. Our tests used $\epsilon = 1e^{-8}$.

[4] `https://github.com/Alibaba-MIIL/ASL`

Table 1. Results on the VQA-MED 2019 dataset for MMBERT models pre-trained with masked language modeling. The highlighted row reports the result for our run with the default hyper-parameters in the original source code.

Batch Size	Patience	Stop Criteria	Accuracy	BLEU
MMBERT	-	-	62.4	64.20
Reproduced results				
16	40	Loss	48.2	50.62
48	40	Loss	55.6	57.65
48	20	Loss	56.0	58.36
16	**20**	**Accuracy**	**58.8**	**60.74**
16	40	Accuracy	58.9	60.88
48	20	Accuracy	59.6	61.36
48	40	Accuracy	59.8	61.59
48	80	Accuracy	59.2	61.27

5 Experimental Evaluation

We tested the original MMBERT model, as well as our extensions, on the ImageCLEF 2019 VQA-MED dataset [3]. The dataset contains 12792 question-answer pairs and 3200 medical images for training, together with 2000 question-answer pairs and 500 images for validation. There are also 500 question-answer pairs and 500 images for testing. Each question-answer pair belongs to one of 4 different categories: modality, plane, organ system, and abnormality. In the training split, the number of different possible answers for each category is as follows: 44 for modality; 15 for plane; 10 for organ system; and 1485 for abnormality. In the validation split, there are 149 questions (29.4%) whose answers do not exist on the training split. Since we are addressing the VQA problem as a classification task, these are answers that the model will never be able to produce. These correspond to 11% of the total possible answers in test split.

VQA-MED used two primary evaluation metrics, namely strict accuracy and BLEU [22]. Strict accuracy considers the exact matching of a provided answer and the ground-truth answer, corresponding to the percentage of correct classifications. BLEU is not as strict, and is used to capture the word overlap similarity between a system-generated answer and the ground-truth answer. While BLEU is more commonly used for the evaluation of text generation, the ImageCLEF organizers considered BLEU even when treating VQA as a classification task, mostly due to the hierarchical relationship of the candidate answers. For example, given a ground-truth answer `ct scan with contrast` and a prediction of `ct scan`, the strict accuracy would evaluate to 0 but BLEU would consider the fact that the system got two words correctly.

The chosen hyper-parameters follow the original MMBERT source code. For pretraining and fine-tuning, images are resized to 224×224 pixels. To avoid distortion, first the smaller edge is matched to 224, and then a center crop of size 224×224 is performed. Image crops, rotations, and color jittering are used for augmentation. For pre-training, the optimizer is Adam and the learning rate is $2e^{-5}$, which is reduced if the validation loss does not improve for 5 consecutive

Table 2. Results on VQA-MED 2019 dataset for different configurations and pre-training objectives. The first row is the original architecture of MMBERT. SupCon-J represents the use of the SupCon loss when computing the similarity between sentences with the Jaccard similarity. SupCon-SB represents the SupCon loss with the cosine similarity between the sentence-BERT encodings.

Image Encoder	Architecture	Activation	Pretraining task	Loss	Accuracy	BLEU
ResNet152	Transformer	ReLU	MLM	CE	58.80	60.74
Effic.NetV2	Transformer	ReLU	MLM	CE	59.40	61.36
Effic.NetV2	RealFormer	SERF	MLM	ASL	59.80	61.55
Effic.NetV2	RealFormer	SERF	MLM + SimCLR	ASL	59.80	61.5
Effic.NetV2	RealFormer	SERF	MLM + SupCon-J	ASL	60.20	62.5
Effic.NetV2	RealFormer	SERF	MLM + SupCon-SB	ASL	60.60	62.98
Effic.NetV2	RealFormer	SERF	MLM + SupCon-SB	ASL	61.60†	63.72†
Effic.NetV2	RealFormer	SERF	MLM + SupCon-SB	ASL	62.80†*	64.32†*

*† represents a model where the batch size was set to 48 (vs 16 in the rest), and * represents a model where the patience was set to 80.*

epochs. As per the source code, models are pre-trained for 10 epochs and the probability of masking a given medical token is 0.15. For fine-tuning, the Adam optimizer is also used, with a learning rate of $1e^{-4}$ which is reduced by a factor of 0.1 if validation loss does not improve for 10 consecutive epochs. The batch size was set to 16, but we also tested the use of larger batches with 48 instances. We use the loss or the accuracy in the validation split as early stopping criteria, with different values for patience.

The reproduction of the results reported for the original model was met with some difficulties, given that the default hyper-parameters present in the original source code resulted in a lower performance. Table 1 shows these results, where the 4-th row is a reproduction with these hyper-parameters. Using a larger batch size or training for more epochs (i.e., using a higher value for patience) helps to increase the performance, but the original results of MMBERT were never met. Although the MMBERT paper does not mention the values for the aformentioned hyper-parameters, the authors claim to be using a NVIDIA RTX 2080Ti GPU with 11 GB of memory, and hence the batch size used in their tests was likely less than 48.

Table 2 shows results for the proposed extensions, also assessing the performance of different pre-training tasks based on contrastive learning. Compared to a model with the same architecture but employing only the MLM objective, the use of contrastive learning benefits the performance. All but the last rows in Table 2 used the same hyper-parameters given in the original source code but,

Table 3. Results per category on the VQA-MED 2019 dataset, for two different configurations. MMBERT represents the results for the original architecture and default hyper-parameters. MMBERT-SupCon-SB†* is detailed in the last row of Table 2.

Method	Modality		Plane		Organ		Abnormality		Binary		Overal	
	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU
MMBERT	70.8	77.2	80.8	80.8	68.8	71.8	0.9	2.4	81.3	81.3	58.2	60.2
MMBERT-SupCon-SB†*	76.4	81.2	80.8	80.8	72.0	74.8	13.2	13.8	82.8	82.8	62.8	64.3

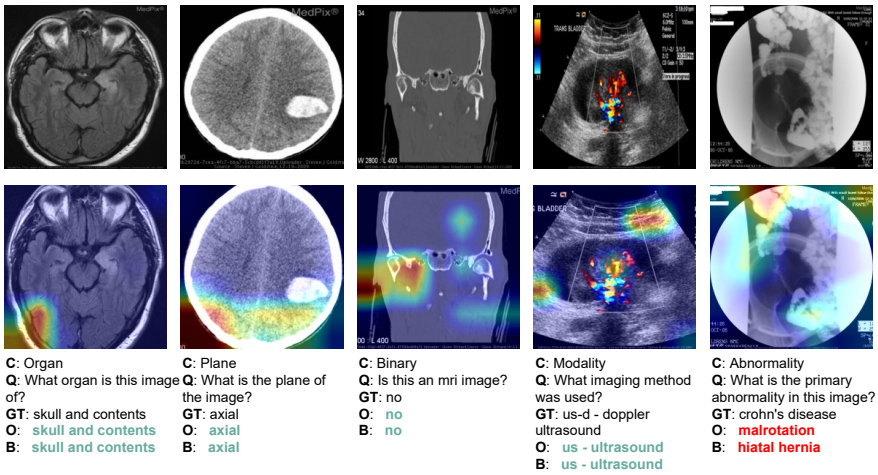

Fig. 2. Graphical depiction of Grad-cam activation maps. **B** corresponds to MMBERT, GT to ground-truth, and **O** corresponds to the model marked with † in Table 2.

as suggested by various studies on contrastive learning, increasing the batch size leads to a higher performance, as did training the model with more epochs.

Table 3 presents results for different question categories. Our reproduction of the original MMBERT architecture had poor results in the abnormality category, while getting reasonably high results on the other categories. The full model is able to improve results over all categories, and approximately by 13 times in the abnormality category, although this result is still significantly lower.

Figure 2 illustrates the results obtained with the original MMBERT and with our extensions. The top row shows the original images, and the second row shows the results of a method named gradient-weighted class activation map (Grad-CAM) [27] to assess the contribution of different image regions. The gradients of the final EfficientNetV2 block were used to construct the localization map emphasizing the important regions for a specific classification. In the organ, plane, and binary categories, the model attends the bony and tissue parts surrounding the skull, as well as the brain, to produce an answer. For the abnormality example, both methods fail to retrieve the correct answer. Still, our method provides an answer with an issue related to the organ of the ground truth (intestines), while the answer of the simpler model is related with stomach issues.

6 Conclusions and Future Work

This work attempted to reproduce the results of a recent multimodal Transformer model for visual question answering in the medical domain, leveraging pre-training with in-domain data. We had some difficulties in reproducing the originally reported results, and values can change considerably according to hyper-parameter choices. We also assessed a number of possible extensions to the original model, showing that they can lead to improvements. For future work, we can consider additional extensions and/or model changes. These include other methods for visual token extraction [1], or other Transformer extensions [28].

References

1. Bao, H., Dong, L., Wei, F.: BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 (2021)
2. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: Overview of the VQA-MED Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. In: Proceedings of the Cross Language Evaluation Forum (2020)
3. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-MED: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: Proceedings of the Cross Language Evaluation Forum (2019)
4. Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric Loss For Multi-Label Classification. arXiv:2009.14119 (2021)
5. Ben-younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: Multimodal Tucker Fusion for Visual Question Answering. arXiv:1705.06676 (2017)
6. Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D.: Multimodal Pretraining Unmasked: Unifying the Vision and Language BERTs. arXiv:2011.15124 (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 (2020)
8. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. arXiv:1909.11740 (2019)
9. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying Vision-and-Language Tasks via Text Generation. arXiv:2102.02779 (2021)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2019)
11. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. arXiv:1606.01847 (2016)
12. Gupta, S., Akin, B.: Accelerator-aware Neural Network Design using AutoML. arXiv:2003.02838 (2020)
13. He, R., Ravula, A., Kanagal, B., Ainslie, J.: RealFormer: Transformer Likes Residual Attention. arXiv:2012.11747 (2020)
14. Hoang Minh, V., Sznitman, R., Nyholm, T., Löfstedt, T.: Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain. In: Proceedings of the Cross Language Evaluation Forum (2019)
15. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-Excitation Networks. arXiv:1709.01507 (2019)
16. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.V.: MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1033–1036 (2021)
17. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised Contrastive Learning. arXiv:2004.11362 (2021)
18. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-Thought Vectors. arXiv:1506.06726 (2015)
19. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. arXiv:1908.02265 (2019)

20. Nag, S., Bhattacharyya, M.: SERF: Towards better training of deep neural networks using log-Softplus ERror activation Function. arXiv:2108.09598 (2021)
21. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the International Conference on Machine Learning (2010)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2002)
23. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: Stoyanov, D., Taylor, Z., Balocco, S., Sznitman, R., Martel, A., Maier-Hein, L., Duong, L., Zahnd, G., Demirci, S., Albarqouni, S., Lee, S.L., Moriconi, S., Cheplygina, V., Mateus, D., Trucco, E., Granger, E., Jannin, P. (eds.) Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. pp. 180–189. Springer International Publishing (2018)
24. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 (2019)
25. Ren, F., Zhou, Y.: CGMVQA A New Classification and Generative Model for Medical Visual Question Answering. IEEE Access:10.1109ACCESS.2020.2980024 (2020)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 (2019)
27. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. arXiv:1610.02391 (2019)
28. Shleifer, S., Weston, J., Ott, M.: NormFormer: Improved Transformer Pretraining with Extra Normalization. arXiv:2110.09456 (2021)
29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
30. Srivastava, Y., Murali, V., Dubey, S.R., Mukherjee, S.: Visual Question Answering using Deep Learning: A Survey and Performance Analysis. arXiv:1909.01860 (2020)
31. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (2020)
32. Tan, M., Le, Q.V.: EfficientNetV2 Smaller Models and Faster Training. arXiv:2104.00298 (2021)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762 (2017)
34. Yan, X., Li, L., Xie, C., Xiao, J., Gu, L.: Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain. In: Proceedings of the Cross Language Evaluation Forum (2019)
35. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. arXiv:1708.01471 (2017)