Supplementary Material for
“Multi-task Learning of Hierarchical Vision-Language Representation”

Duy-Kien Nguyen¹ and Takayuki Okatani¹,²

¹Graduate School of Information Sciences, Tohoku University
²RIKEN Center for Advanced Intelligence Project
{kien, okatani}@vision.is.tohoku.ac.jp

This document contains the following: i) more details of setup of the experiments reported in the main paper (Sec.A); ii) additional results of the selection of optimal layers for the task-specific decoders (Sec.B); iii) additional results of ICR on MSCOCO 5,000 testing images (Sec.C); and iv) more visualization of the proposed network for a variety of images including failure cases of VQA and ICR (Sec.D and E).

A. More Details of the Experimental Setup

In all the experiments reported in this study, images and sentences (i.e., questions or captions) were preprocessed as follows. We used Faster-RCNN [9] to extract the bottom-up features from each image, which yields from 10 to 100 features (also referred to as regions in this study), i.e., \( T \in [10, 100] \). Questions and captions were tokenized using Python Natural Language Toolkit (nltk) [2]. We used the vocabulary provided by the CommonCrawl-840B GloVe model for English word vectors [8], and set out-of-vocabulary words to \( \text{unk} \).

As is mentioned in Sec.4.2 of the main paper, we conducted a hyper-parameter search on several training parameters including the layers used for the task-specific decoders by training the network on each individual task. The parameters thus determined are shown in Table 1.

We provide below additional details of the training procedures of the three tasks. We used the cross-entropy loss for all the three tasks.

**Image Caption Retrieval** This task consists of two subtasks; one is to retrieve relevant images given a query caption (image retrieval) and the other is to retrieve relevant captions given a query image (image annotation). In the training, given pairs of image-caption \((I, C)\)’s, where \(I\) and \(C\) are an image and a caption, respectively, we compute the losses for the two subtasks for each pair as follows. For image retrieval, we randomly sample \(F - 1\) images that are different from \(I\), and compute the loss for \(F\) images including \(I\), in which the label for the ground truth image \((I)\) is set to 1 and those for the others are all 0.

<table>
<thead>
<tr>
<th>Task</th>
<th>Level</th>
<th># step</th>
<th># iter</th>
<th>Batch size</th>
<th>Cycle (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA</td>
<td>5</td>
<td>8K</td>
<td>20K</td>
<td>400</td>
<td>1</td>
</tr>
<tr>
<td>ICR</td>
<td>3</td>
<td>12K</td>
<td>30K</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>VG</td>
<td>2</td>
<td>4K</td>
<td>10K</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>VQA + VG</td>
<td>5, 2</td>
<td>12K</td>
<td>30K</td>
<td>400, 64</td>
<td>3</td>
</tr>
<tr>
<td>ICR + VG</td>
<td>3, 2</td>
<td>16K</td>
<td>40K</td>
<td>64, 64</td>
<td>4</td>
</tr>
<tr>
<td>VQA + ICR</td>
<td>5, 3</td>
<td>20K</td>
<td>50K</td>
<td>400, 64</td>
<td>5</td>
</tr>
<tr>
<td>VQA + ICR + VG</td>
<td>5, 3, 2</td>
<td>24K</td>
<td>60K</td>
<td>400, 64, 64</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: Hyperparameters determined by training on individual tasks and then used for joint training (# step: step size of learning rate decay, # iter: total of training iterations, K=1,000 units).
For image annotation, we randomly sample \( F - 1 \) captions that are not the ground truth captions corresponding to the image \((I)\) and compute the loss of image annotation for \( F \) captions including \( C \), in which the label for the ground truth caption \((C)\) is 1 and those for the others are 0. We minimize the sum of the two losses. We used \( F = 26 \) for all the experiments.

**Visual Question Answering**  We followed the procedure of [9]. We treat VQA as a multi-label classification task, where each training question is associated with one or several answers with soft accuracy label(s) in [0, 1]. Multiple answers appear in the case of disagreement among human annotators. The scores of answers are computed as in the original paper [1], that is,

\[
\text{score}_i = \min \left( \frac{\# \text{humans that provide the } i\text{-th answer}}{3}, 1 \right)
\]

where \( \text{score}_i \) is the score of \( i\)-th answer in the predefined answer set.

**Visual Grounding**  The dataset provides a set of samples, each of which is built upon a pair of an image and its caption. Each sample consists of a set of phrases in the caption and the corresponding box(es) in the image. We label each phrase with its corresponding box(es) as 1 and with other boxes as 0. These boxes are obtained in the aforementioned pre-processing using the pre-trained Faster-RCNN. The loss is the sum over all possible phrase-region pairs in the image and caption.

**B. Additional Results of Layer Selection for the Three Tasks**

As mentioned in Sec.5.2 of the main paper, we train our network on each individual task to choose the layer of the shared encoder fit for each task. Table 2 shows the results. Based on these, we determined \( l_R = 3 \) (image caption retrieval), \( l_Q = 5 \) (VQA), and \( l_G = 2 \) (visual grounding), as reported in the main paper. However, while it is simple and can be performed efficiently, this method may not provide optimal choice of layers for the three tasks, as it does not consider interactions among the three tasks.

Table 2: Performance of the proposed network trained and tested on the same individual task. These are used to determine the layer of the shared encoder for each of the three tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Layer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA</td>
<td></td>
<td>64.72</td>
<td>65.21</td>
<td>65.34</td>
<td>65.35</td>
<td><strong>65.50</strong></td>
<td>65.27</td>
</tr>
<tr>
<td>ICR</td>
<td>56.45</td>
<td>56.78</td>
<td><strong>57.15</strong></td>
<td>54.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VG</td>
<td>57.74</td>
<td><strong>58.09</strong></td>
<td>57.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Thus, we also tested another method for choosing the layers that is based on joint-training of the three tasks. Table 3 shows the results, which were obtained by the following procedure. Initially, we determine the order of the three tasks in terms of the level in the hierarchy of the shared representation. Based on the above results, we determine their (descending) order as follows: VQA, ICR, and VG. We first determine the optimal layer for VQA by training the network on VQA alone, which is the same as the first row (VQA) of Table 2; this results in \( l_Q = 5 \). Next, we determine the optimal layer for ICR. To do this, we train the network on VQA+ICR for different choice of the layer for ICR (i.e., \( l_R = 1, \ldots \)) while fixing the layer for VQA (i.e., \( l_Q = 5 \)). We evaluate the performance for different \( l_R \)’s on VQA, ICR (image annotation) and ICR (image retrieval). The second to forth rows of Table 3 show the performance on VQA, image annotation, and image retrieval, respectively. From this, we choose \( l_R = 3 \). Finally, we determine the layer for VG. To do this, we train the network on VQA+ICR+VG for different layer \( l_G \) (= 1, 2, 3) for VG. As above, we evaluate the performance on VQA, image annotation, and image retrieval, and VG, which are shown in the fifth to eighth rows of Table 3, respectively. From this, we choose \( l_G = 2 \).

In short, we obtain the same results as the first method based on individual task training. This confirms the validity of our choice of the layers for the three tasks. In the above experiments, we set \( F = 16 \) in image caption retrieval for efficient computation; the reduction of \( F \) contributes the most to reducing necessary computational resource.
Table 3: Performance of the proposed network trained and tested on several combinations of tasks. Their combinations are created in a cumulative fashion, assuming the order of the three tasks to be VQA, ICR, and VG in terms of level in the representation hierarchy in the shared encoder. ICR1 and ICR2 indicate image annotation and image retrieval, respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>Layer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained/tested on VQA alone</td>
<td>(VQA)</td>
<td>64.72</td>
<td>65.21</td>
<td>65.34</td>
<td>65.35</td>
<td>65.50</td>
<td>65.27</td>
</tr>
<tr>
<td>Trained/tested on VQA+ICR(2 subtasks) for different layers for ICR and layer = 5 for VQA</td>
<td>(VQA)</td>
<td>65.80</td>
<td>66.00</td>
<td><strong>66.09</strong></td>
<td>65.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(ICR1)</td>
<td>58.01</td>
<td>58.63</td>
<td><strong>59.25</strong></td>
<td>56.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(ICR2)</td>
<td>47.91</td>
<td>48.70</td>
<td><strong>49.03</strong></td>
<td>47.27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Trained/tested on VQA+ICR(2)+VG for different layers for VG and layer = 5 for VQA and 3 for ICR</td>
<td>(VQA)</td>
<td>66.10</td>
<td>66.15</td>
<td><strong>66.15</strong></td>
<td>66.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(ICR1)</td>
<td><strong>61.94</strong></td>
<td>61.88</td>
<td>61.90</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(ICR2)</td>
<td>49.84</td>
<td>50.81</td>
<td><strong>50.81</strong></td>
<td>50.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(VG)</td>
<td>57.86</td>
<td><strong>58.17</strong></td>
<td>57.83</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

C. Comparisons on MS-COCO dataset of 5,000 testing images

As noted in the main paper, we conducted evaluation on MSCOCO 5,000 testing images. The results are shown in Table 4. Our method is comparable to the state-of-the-art method (S-E Model). It is noteworthy that our method provided only 50 mismatched pairs for each matched pair, while S-E Model provided 128.

Table 4: Results of image annotation and retrieval on the MSCOCO (5,000 testing) datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Annotation</th>
<th>Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>DVSA [5]</td>
<td>11.8</td>
<td>32.5</td>
</tr>
<tr>
<td>FV [6]</td>
<td>17.3</td>
<td>39.0</td>
</tr>
<tr>
<td>OEM [10]</td>
<td>23.3</td>
<td>50.5</td>
</tr>
<tr>
<td>VQA [7]</td>
<td>23.5</td>
<td>50.7</td>
</tr>
<tr>
<td>VSE++ [3]</td>
<td>41.3</td>
<td>69.2</td>
</tr>
<tr>
<td>S-E Model [4]</td>
<td><strong>42.8</strong></td>
<td><strong>72.3</strong></td>
</tr>
<tr>
<td>Ours</td>
<td>42.2</td>
<td>69.1</td>
</tr>
</tbody>
</table>
D. Visualization of Inference of Three Tasks on Visual Question Answering

In Fig. 2 of the main paper, we show a few examples of visualization of inference of VG, ICR, and VQA on complementary image-question pairs of VQA 2.0 (i.e., pairs of the same questions and different images). We show here more examples for success cases (Sec. D.1) as well as failure cases (Sec. D.2).

D.1. Success Cases

We first show visualization for success cases, i.e., image-question pairs for which our network provides the correct answers. As in Fig. 2 of the main paper, each of the left and right panels on each page shows visualization for a complementary image-question pair. The same observation given in the main paper applies to these examples.
is the woman standing

the women standing
matching score: 1.225e-05

is the woman standing

the women standing
matching score: 0.007

is the horse jumping

the horse jumping
matching score: 0.825

is the horse jumping

the horse jumping
matching score: 0.339

is the women standing
Pred: no, Ans: no

is the women standing
Pred: yes, Ans: yes

is the horse jumping
Pred: yes, Ans: yes

is the horse jumping
Pred: no, Ans: no
what is the man doing
Pred: talking on phone, Ans: talking on phone

what is the man doing
Pred: drinking, Ans: drinking

what color is the man's pants
Pred: green, Ans: green

what color is the man's pants
Pred: blue, Ans: blue
how many horses are in the picture

how many different poses are in this shot

how many horses are in the picture

how many different poses are in this shot

how many horses are in the picture

how many different poses are in this shot

how many horses are in the picture

how many different poses are in this shot

how many horses are in the picture

how many different poses are in this shot
how many pictures are on the wall

how many pictures are on the wall

what color is lit up on the street lights

what color is lit up on the street lights

how many pictures are on the wall

how many pictures are on the wall

what color is lit up on the street lights

what color is lit up on the street lights

Pred: 0, Ans: 0

Pred: 2, Ans: 2

Pred: green, Ans: green

Pred: white, Ans: white
D.2. Failure Cases

We next show failure cases, i.e., image-question pairs for which our network provides at least one wrong answer for the VQA task. The red bounding boxes indicate wrong answers and the green ones indicate correct answers. From the examples shown below, we can categorize failures for the VQA task into the following typical cases, for each of which we can explain why our network provides wrong answers and suggest possible solutions:

1) Although the VG and ICR decoders are able to correctly locate objects or concepts in the input image that appear in the input question, the VQA decoder fails to distinguish different objects or concepts that have similar appearance. This may be attributable to that the pretrained Faster R-CNN used for extracting image features is not trained to distinguish fine-grained concepts (e.g., “terrier” and “lab” (i.e., Labrador retriever), “round” and “oval”). It may help to train the Faster R-CNN with more fine-grained concepts.

2) The network is unable to locate relevant image regions. This often occurs when the VG and ICR decoder at the lower layer of the network are unable to detect correct regions (e.g., “the bottom corner” or “inside of the plane”), which leads to the failure of the VQA decoder. This is mostly because the Faster R-CNN fails to extract right regions or extracts excessively large regions containing many objects.

3) Questions require general knowledge that cannot be learned from only the training data (e.g., “acidic food” or “reflection”). For instance, for the question “is this acidic food”, we can observe from the response of the VG and ICR decoders that the network recognizes all the food in the image as “acidic food”; for the question “is there a reflection in the window”, the VG and ICR decoders give high confident scores for “reflection” even there is not.

4) The answers given by the network are judged incorrect simply because they are not listed in the given set of correct answers in the dataset, but they are actually considered to be correct answers. For example, for the question “what is on the floor by the toilet”, both of “tile” and “trash can” should be correct answers, but only the latter is listed in the correct answer set.
is there a reflection in the window

what is the number in orange on the white shirt

what is the number in orange on the white shirt

what is the number in orange on the white shirt
is there a pilot inside of the plane

what color is she wearing

Pred: no, Ans: no

Pred: no, Ans: yes

Pred: black, Ans: orange

Pred: gray, Ans: gray
what [branch] [bat] is [the boy] in [the blue] [shirt] using
what is on [the floor] by [the toilet]
what brand bat is the boy in the blue shirt using
what brand bat is the boy in the blue shirt using
what is on the floor by the toilet
what is on the floor by the toilet

Pred: wilson, Ans: nike
Pred: wilson, Ans: wilson
Pred: tile, Ans: trash can
Pred: toilet brush, Ans: toilet brush
what [year] does it say in [the bottom] [corner]

what [shape] is [the coffee] [table] in [the living] [room]

what year does it say in the bottom corner

what shape is the coffee table in the living room

what year does it say in the bottom corner

date: 0, Ans: 2007

what year does it say in the bottom corner

date: 2000, Ans: 2010

what shape is the coffee table in the living room

Pred: round, Ans: oval

what shape is the coffee table in the living room

Pred: rectangle, Ans: rectangle
what color are the curtains
Pred: white, Ans: white

what color are the curtains
Pred: white, Ans: yellow and white

what breed of dog is this
Pred: terrier, Ans: lab

what breed of dog is this
Pred: terrier, Ans: terrier
E. Visualization of Inference of VG and ICR on Image Caption Retrieval

We show visualization of our network using an ICR dataset, MSCOCO. As success cases are not so informative, which are often shown in previous studies, we show only failure cases for the two subtasks of ICR, more specifically, the cases where the top-1 prediction is not correct for image annotation (Sec. E.1) and image retrieval (Sec. E.2). Each panel consists of $2 \times 2$ image-caption pairs; the first row shows VG and the second row shows ICR; the first column (with a green box) shows VG and ICR visualization for the ground-truth image-caption pair and the second column (with a red box) shows a pair of an input image and the predicted top-1 caption for image annotation and a pair of the predicted top-1 image and an input caption for image retrieval.

We think that the failure cases can be categorized into the following three types:

1) The network incorrectly recognizes objects or concepts that have similar appearance, since the Faster R-CNN features are not rich enough to distinguish them (e.g., “computer monitor” and “television”). Specifically, the VG decoder detects wrong objects, matching the input image with a wrong caption.

2) The VG and ICR decoders correctly align objects in the input caption with the corresponding image regions, but fail to recognize their actions (e.g., “submerged in a small body of water” vs. “standing on a rock”; “perches” vs. “standing”). Such failures may be eliminated by creating a dataset for the task of predicting such actions and using it in the joint training.

3) As in the case of VQA, although the top-1 caption or image predicted by our network matches well with the input image or caption, it is judged wrong because it is not listed in the set of correct answers. For image annotation, this happens because the same image content can be described in many ways (e.g., “a train on a track near a field with tall grass” or “colorful train cars are on the track next to some grass”); or because “correct” captions explain only one of multiple contents contained in input images (e.g., “a man on the beach who is carrying a surfboard” and “multiple people are standing on the beach at the edge of the water”). For image retrieval, the same often occurs when the input captions provide only too general explanation of a scene, most of which tend to be short simple captions, such as “there are people flying kites in the park”. It should be noted that even if this is the case, the VG decoder is able to correctly detect objects in most cases.
E.1. Image Annotation

[a train] on [a track] near [a field] with [tall grass]

[family] of [animals] in [a zoo]

[a man] on [the beach] who is carrying [a surfboard]

[a man] with [a helmet] on [a motorcycle] pulled over

[a man] posing with [a motorcycle] on [a dry terrain]

[colorful train] cars are on [the track] next to [some grass]

[family] of [animals] in [a zoo]

[a man] on [the beach] who is carrying [a surfboard]

[a man] with [a helmet] on [a motorcycle] pulled over

[a man] posing with [a motorcycle] on [a dry terrain]

[a woman] and [a young girl] sit on [a bed] eating bananas

[a woman] that is sitting down with [a baby]
[a polar bear] fully submerged in a small body of [water]
[a large white bear] standing on [a rock]
[a man] sitting down with [an angry look]
[a man] who is wearing [a tie] that is too small for him

[a snowboarder] does [a trick] in [front] of [a building]
[a person] doing [a snowboarding] [stunt] on [a rail]
[an orange tabby] [cat] perches near [a computer] [monitor]
[a cat] is sitting on [the floor] and watching [television]
[a person] riding [a surf board] on [a wave]

[a surfer] rides [a wave] while [another surfer] paddles towards [the wave]

a close up of [a person] playing [a video] [game]

[a man] with [long hair] gets a [hair cut]

[a person riding a surf board on a wave]
a surfer rides a wave while another surfer paddles towards the wave

[a close up] playing a video game

[a man with long hair getting a haircut]

[young boy] sitting up in [his crib] playing with [clock] [radio]

[a baby] that is pressing on a small, [black suitcase]

[a plate] of [food] that is on [a table]

there is some [tuna] and [a potato] on [a white plate]

young boy sitting up in his crib playing with clock radio

baby that is pressing on a small black suitcase

a plate of food that is on a table

there is some tuna and a potato on a white plate
[an orange tabby cat] perches near [a computer] [monitor]

[a dude] on [a bike] rides quietly across [the place]

[a person] riding [a surf board] on [a wave]

there are people flying kites in [the park]
this is some animals sitting in [the dirt]

[a little girl] doing some arts and crafts

[a smiling man] stands on [a snowy hill] with [some ski] poles

[a man] and [woman] look into each others eyes while getting married

[a man and woman] look into each others eyes while getting married

[a man and woman] look into each others eyes while getting married

[a man and woman] look into each others eyes while getting married

[a smiling man] stands on a snowy hill with some ski poles

[a smiling man] stands on a snowy hill with some ski poles
References


