Free VQA Models from Knowledge Inertia by Pairwise Inconformity Learning

Yiyi Zhou,[1] **Rongrong Ji,**[1,3] **Jinsong Su,**[2] **Xiangming Li,**[1] **Xiaoshuai Sun**[1]*
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China
[2]School of Software Engineering, Xiamen University, China
[3]Peng Cheng Laboratory, China

Abstract

In this paper, we uncover the issue of *knowledge inertia* in visual question answering (VQA), which commonly exists in most VQA models and forces the models to mainly rely on the question content to "guess" answer, without regard to the visual information. Such an issue not only impairs the performance of VQA models, but also greatly reduces the credibility of the answer prediction. To this end, simply highlighting the visual features in the model is undoable, since the prediction is built upon the joint modeling of two modalities and largely influenced by the data distribution. In this paper, we propose a Pairwise Inconformity Learning (PIL) to tackle the issue of knowledge inertia. In particular, PIL takes full advantage of the similar image pairs with diverse answers to an identical question provided in VQA2.0 dataset. It builds a multi-modal embedding space to project pos./neg. feature pairs, upon which word vectors of answers are modeled as anchors. By doing so, PIL strengthens the importance of visual features in prediction with a novel *dynamic-margin based triplet loss* that efficiently increases the semantic discrepancies between pos./neg. image pairs. To verify the proposed PIL, we plug it on a baseline VQA model as well as a set of recent VQA models, and conduct extensive experiments on two benchmark datasets, *i.e.*, VQA1.0 and VQA2.0. Experimental results show that PIL can boost the accuracy of the existing VQA models (1.56%-2.93% gain) with a negligible increase in parameters (0.85%-5.4% parameters). Qualitative results also reveal the elimination of knowledge inertia in the existing VQA models after implementing our PIL.

Introduction

Visual question answering (VQA) is a task of answering the human question based on a given image. The challenge lies in how to jointly understand comprehensive visual and textual information. The existing VQA models (Lu et al. 2016; Fukui et al. 2016; Yu et al. 2017; Anderson et al. 2018) have demonstrated the ability to answer human questions in various tasks, such as text understanding, object or scene recognition, counting and visual reasoning *etc*.

Despite the exciting progress, the strong language priors has long plagued the development of VQA (Antol et al. 2015; Goyal et al. 2017). In particular, as revealed latter in Sec.3, the existing VQA models tend to rely solely on the content of a given question to predict the answer, without regard to the given image, *i.e.*, a phenomenon termed as *knowledge inertia* (Liao 2002)[1] in this paper. Fig.1.(a) illustrates two typical cases of knowledge inertia.

To eliminate the negative effect of knowledge inertia, a natural solution is to force the VQA model to make predictions more based on the visual content, *i.e.*, by emphasizing the weightings of visual features, as argued in Goyal et al. (2017). However, doing so is intractable in the existing VQA models. First, the objective function of existing VQA models is to maximize the answer probability conditioned on the given question and image. The prediction is built upon the joint modeling of the textual and visual features, where the weightings of two modalities are automatically learned by the training examples and can not be manually adjusted. Second, as a classification task, VQA also faces the problem of extremely uneven data distributions, which subsequently exacerbates the issue of knowledge inertia.

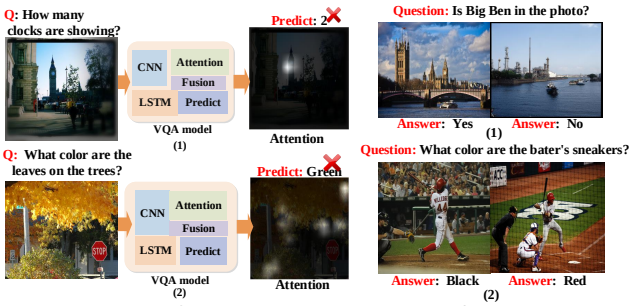


Figure 1: Illustrations of the knowledge inertia (a) and the image pair setting (b). In (a), the model focuses on the correct answer entities in images, but still predicts the wrong answer due to the negative impact of knowledge inertia. In (b), the pair of similar images has the same question, similar content but different answers. This setting is used in this paper to enhance the role of visual information in prediction.

*Corresponding Author. E-mail: xiaoshuaisun.hit@gmail.com

[1]It is borrowed from social science to define a solving strategy that uses redundant and stagnant knowledge and past experience without regard to new inputs.

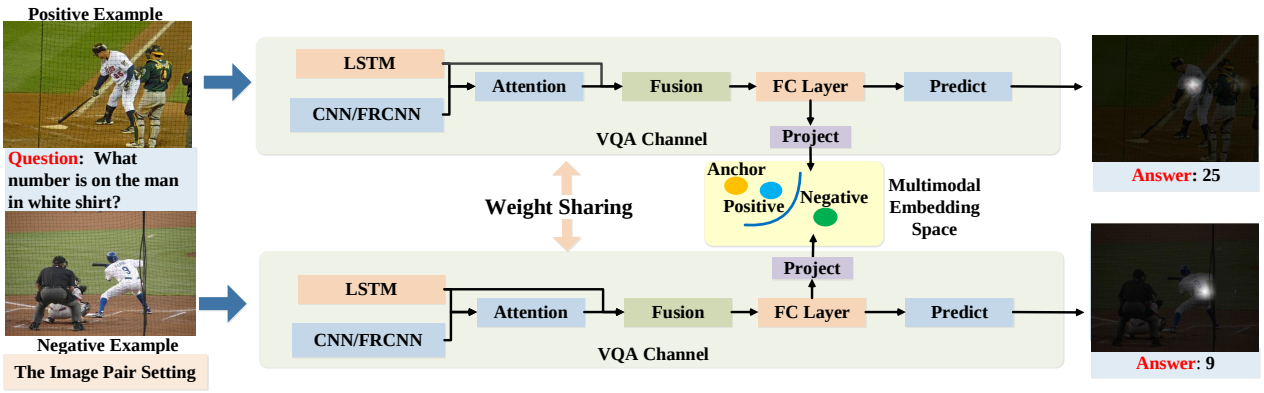


Figure 2: Framework of Pairwise Inconformity Learning (PIL). The input is a pair of similar images with the identical question but different answers. The model processes this pair of examples simultaneously, and projects their joint features onto a multi-modal embedding space. The blue and green dots refer to the embedded joint features of the positive and negative examples, while the orange dot denotes the anchor which is the pre-trained word vector of the positive example's answer.

In this paper, we propose a novel *Pairwise Inconformity Learning*, denoted as PIL[2], which takes full advantage of the image pair setting proposed in VQA2.0 dataset (Goyal et al. 2017), an example of which is shown in Fig.1.(b). In principle, under the same textual input, we target at maximizing the semantic distances between joint features of positive and negative examples, which subsequently eliminates the impact of knowledge inertia in prediction. In particular, PIL first creates a multi-modal embedding space in addition to the typical VQA model settings. In this space, a novel *dynamic margin based triplet loss* is further proposed to increase the semantic distance between pos./neg. examples, where the well pre-trained word embedding of the answer is used as anchor. Then, hard example mining is performed based on the pairwise distances, which further improves the VQA performance.

The proposed PIL is flexible, which can be well plugged into a range of standard and cutting-edge VQA models, such as SAN (Yang et al. 2016) and BUA (Teney et al. 2018). We have quantified its accuracy boost by evaluating it on VQA1.0 (Antol et al. 2015) and VQA2.0 (Goyal et al. 2017) datasets, which reports 1.6%-2.9% gains over two datasets. Moreover, the increase of model complexity is limited, say an increase of 0.85%-5.4% in parameters, which retains highly efficient in real-world applications.

Overall, the contribution of this paper is three-fold:

- We uncover a key issue in VQA, namely *Knowledge Inertia*, which restricts the performance and the credibility of the existing VQA models.
- We propose a Pairwise Inconformity Learning (PIL) scheme to address this issue, which includes a dynamic margin based triplet loss and an online hard example mining strategy to boost the VQA performance.
- PIL is generalized and can boost the accuracy of the existing VQA models (1.2%-3.5% gain) with a negligible increase in parameters (0.85%-3.7% parameters).

[2]https://github.com/xiangmingLi/PIL

Related Work

At present, visual question answering (VQA) is often considered as a classification task with fixed categories (Antol et al. 2015; Yang et al. 2016; Fukui et al. 2016; Anderson et al. 2018). Earlier, VQA methods tend to predict the answer by directly learning a joint representation of the image and text features extracted by convolutional neural network (CNN) and recurrent neural networks (RNNs) (Antol et al. 2015)(Ma, Lu, and Li 2015). To capture the most relevant visual signal, the work in (Yang et al. 2016) first introduces the attention mechanism to VQA, which is further extended to a multi-step attention to improve visual reasoning. There are also some recent works focusing on the modifications of attention mechanism (Lu et al. 2016; Zhu et al. 2017; Fukui et al. 2016). For example, the work in (Lu et al. 2016) proposes two co-attention algorithms to generate attended features from both visual and textual features. Some recent developments (Fukui et al. 2016; Kim et al. 2017; Yu et al. 2017) focus on investigating different fusion approaches for more efficient interactions between two modalities. For example, the work in (Fukui et al. 2016) uses a compact bilinear pooling for feature fusions, which brings a significant improvement to the model performance. After that, other bilinear pooling based methods are further proposed, such as the work in (Kim et al. 2017; Yang et al. 2016). There are also some works introducing the external knowledge or existing techniques to VQA, such as the use of Wiki knowledge (Wu et al. 2016) and knowledge graph (Wang et al. 2016), or replacing the convolutional feature map with the regional features from Faster-RCNN as the visual input (Anderson et al. 2018).

A long-standing problem that remains unresolved is the existence of strong textual priors in existing datasets, such as the widely-used VQA1.0 dataset (Antol et al. 2015). The most recent VQA dataset, VQA2.0 (Goyal et al. 2017), particularly compensate this defect by balancing the answer distribution and provides a similar image-pair setting. Another recent work (Agrawal et al. 2018) also creates an even

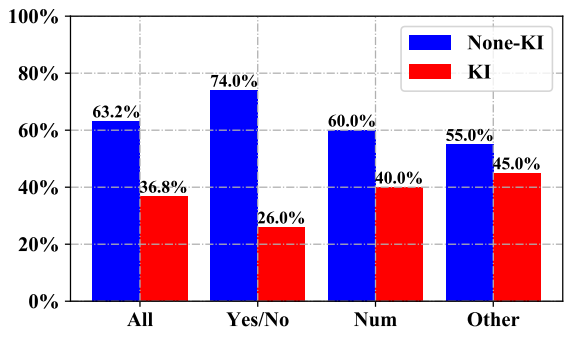


Figure 3: The failure case distribution of 1,000 examples from the VQA2.0 *val* set. "Yes/No", "Num." and "Other" denotes the three main question types in VQA, while "All" represent the overall result. "KI" refers to the failure predictions directly or indirectly caused by Knowledge Inertia, which are the cases that the model finds out the correct answer entities in images but still make false predictions.

dataset based on VQA1.0 and VQA2.0, and propose a model that use visual and textual concepts for answer prediction to avoid the effect of language priors.

Knowledge Inertia

Knowledge inertia is a terminology from social science, which refers to a problem solving strategy uses old, redundant, stagnant knowledge and past experience without recourse to new knowledge and experience (Liao 2002). Here, we use it to represent the behaviors happened in existing VQA models as below.

Ideally, after training, the VQA model should be able to find the most relevant visual signals for prediction based on the question. However, due to the existence of strong language priors in the datasets (Goyal et al. 2017), the model has a strong dependency on the textual features unconsciously. It leads to a result that the model relies sorely on the question to predict answer without regard to visual information. This is termed as knowledge inertia in this paper.

To study the influence of knowledge inertia, we quantitatively analyzed failure cases of the model predictions, the result of which is shown in Fig.3. Here, we define the failure case of knowledge inertia as the model finds the obvious answer in images and still makes wrong predictions, such as the examples in Fig.1.a. It can be seen that at least about 30% of the false predictions are caused directly or indirectly by knowledge inertia, which thereby indicates the emergency to solve this problem.

The key solution to knowledge inertia is to improve the role of visual feature in predictions, so that the questions can be answered based not only on what the model already "know" but also on what the model are currently "seeing". However, directly improving the visual weighting is undoable, as analyzed earlier in Sec.1.

Pairwise Inconformity Learning

To address the issue of knowledge inertia, we propose a Pairwise Inconformity Learning, termed as *PIL*. In addition to the typical VQA setting, PIL creates a multi-modal embedding space with a dynamic-margin based triplet loss to increase semantic distances between positive/negative examples. The framework is shown in Fig.2.

Problem Setup In VQA, The training example is a triplet denoted as $e = (\mathbf{I}, Q, a)$, where $\mathbf{I}$ is the image, Q is the question and a is the answer. The answer prediction is regarded as a classification task with fixed categories, the objective function of which is formulated as:

$$\begin{aligned} &\max_{G} \log P\left(a|f_j\right), \\ &where \quad f_j = G\left(\mathbf{I}, Q\right). \end{aligned} \tag{1}$$

Here, $G\left(\cdot\right)$ is the main part of a VQA model without the prediction layer, and f_j refers to the joint feature.

To improve the role of visual information in prediction, we make use of the image pair setting proposed in VQA2.0 (Goyal et al. 2017), where positive and negative images have the same question, similar visual content, but different answers, as shown in Fig.1.b. Particularly, given a pair of examples, denoted as (e_p, e_n), we aim at increasing the semantic distance between their joint features:

$$\max_{G} d\left(f_j^p, f_j^n\right), \tag{2}$$

where $d\left(\cdot\right)$ is any distance measurement. The intuition is that with the same textual input, the larger semantic distance between pos./neg. examples will result in more discriminative visual representations, forcing the model to rely more on visual information during the prediction .

Multi-modal Embedding. Following this principle, the first step of PIL is to learn a multi-modal embedding space, where the well pre-trained word vectors of the corresponding answers can be used as anchors. Concretely, given a pair of example (e_p, e_n) and a VQA model $G\left(\cdot\right)$, we first extract their multi-modal joint representations, denoted as $f_j^p, f_j^n \in R^m$. Then, we project them into this space and obtain the embedded features, f_e^p and $f_e^n \in R^n$, by:

$$f_e^p = \sigma\left(\mathbf{W}_{emb} f_j^p\right), \quad f_e^n = \sigma\left(\mathbf{W}_{emb} f_j^n\right), \tag{3}$$

where $\mathbf{W}_{emb} \in R^{n \times m}$ is the projection matrix, and $\sigma\left(\cdot\right)$ is an activation function.

The first objective of PIL is to minimize the distances between the embedding vector and anchors:

$$\mathcal{L}_{emb} = \|f_e^p - f_a^p\|_2^2 + \|f_e^n - f_a^n\|_2^2, \tag{4}$$

where f_a^p and $f_a^n \in R^n$ are anchors of pos./neg. examples. The intuition behind this embedding space is three-fold:

- First, we quantitatively found that directly implementing a strong distance regularization (*e.g.*, Eq.2) on f_j will hinder the overall model performance.
- Second, in many metric learning schemes (Schroff, Kalenichenko, and Philbin 2015)(Hermans, Beyer, and Leibe 2017), computing the anchor of each category is

an indispensable step, which is however computationally expensive in VQA due to the large number of categories, *e.g.*, typically over 3,000 in most benchmarks, each with thousands of examples.

- Third, in this embedding space, well pre-trained word vectors of answers can be used as anchors, *e.g.*, Glove Embeddings (Pennington, Socher, and Manning 2014), which already have strong semantics.

Dynamic-margin based Triplet Loss. To increase the feature distances between embedded joint representations of pos./neg. examples, we propose a dynamic-margin based triplet loss $\mathcal{L}_{triplet}$, which is denoted as:

$$\begin{aligned}\mathcal{L}_{triplet} &= \max\left(0, m + \|f_e^p - f_a^p\|_2^2 - \|f_e^n - f_a^p\|_2^2\right) \\ &+ \max\left(0, m + \|f_e^n - f_a^n\|_2^2 - \|f_e^p - f_a^n\|_2^2\right), \\ &where \quad m = \|f_a^p - f_a^n\|_2^2 .\end{aligned} \tag{5}$$

Eq.5 differs from the traditional triplet loss function (Schroff, Kalenichenko, and Philbin 2015) in that the margin value m is dynamically set by the distance between two answer word vectors. Such design can not only avoid the complex tunning of m, but also represent the semantic margins between different answers more accurately.

The Baseline Model. We further propose a baseline model to deploy the proposed PIL, which is depicted in Fig.2. The input image is processed by CNN or Faster R-CNN to obtain a regional feature matrix $\mathbf{F} \in R^{v \times k}$, while the question feature $f_q \in R^h$ is extracted by an LSTM network. Here, k is the number of image regions, and v and h are dimensions of visual and textual features. We then implement an attention process on $\mathbf{F}$ to obtain a weighted visual feature $f_v \in R^v$, formulated as:

$$\begin{aligned}&f_v = \sum_{i=1}^{k} \alpha_i f_i, \\ &where \quad \alpha_i = Softmax\left(e_i\right), \\ &e_i = Conv\Big(ReLu\big(Conv\big(F_{usion}\left(f_i, f_q\right)\big)\big)\Big),\end{aligned} \tag{6}$$

$Conv\left(\cdot\right)$ is a convolution operation and $F_{usion}\left(\cdot\right)$ is a multi-modal fusion function, *e.g.*, an *MFB pooling* (Yu et al. 2017) used in this paper. After that, we fuse f_q and f_v to obtain the joint feature f_j.

In some widely-used datasets, *e.g.* VQA1.0 (Antol et al. 2015) and VQA2.0 (Goyal et al. 2017), a question is often associated with a set of answers. We thereby follow the setting in (Anderson et al. 2018) to formulate the answer prediction as a multi-label classification, and use the *Sigmoid cross entropy* as the loss function, denoted as:

$$\mathcal{L}_{entropy} = \sum_{i}^{N} y_i \log\left(s_i\right) - \left(1 - y_i\right) \log\left(s_i\right), \tag{7}$$

where $s = Sigmoid\left(\mathbf{W} f_j\right)$, $y_i = 1$ indicates the i-category is in the answer set, and N is the number of categories.

The Overall Loss. The overall objective function of PIL:

$$\min_{G, \mathbf{W}_{emb}} \mathcal{L}_{entropy} + \alpha \mathcal{L}_{triplet} + \beta \mathcal{L}_{emb}, \tag{8}$$

where α and β are hyper parameters tuned by experiments.

Hard Example Mining. Our scheme further enables hard example mining by exploring the semantic distance between positive and negative examples. Following the principle of (Schroff, Kalenichenko, and Philbin 2015), we divide the training examples into three categories which are *easy*, *semi-hard* and *hard*. According to this hardness, we further assign a difficulty score d_s as:

$$d_s = \begin{cases} 0.25 & if \quad d\left(f_e^p, f_e^n\right) > m, \\ 0.5 & if \quad d\left(f_e^p, f_a^p\right) < d\left(f_e^p, f_e^n\right) < m, \\ 1.0 & if \quad d\left(f_e^p, f_e^n\right) < d\left(f_e^p, f_a^p\right), \end{cases} \tag{9}$$

where $d\left(\cdot\right)$ is the $l2-$distance, f_e^p and f_e^n are embedded joint features of pos./neg. examples, respectively. f_a is the anchor and m is the dynamic margin. The score will be multiplied by the overall loss during training. By doing so, the model will receive more gradients from hard examples.

Experiments

To validate the proposed learning scheme, we conduct extensive experiments on two widely-used dataset, namely VQA1.0 (Antol et al. 2015) and VQA2.0 (Goyal et al. 2017), and compare out model with a set of stat-of-the-arts.

Experiment Setup

Datasets. VQA1.0 dataset contains 200,000 natural images from MS-COCO (Chen et al. 2015) with 614,153 human annotated questions in total. Each question has 10 free-response answers. The whole dataset is divided into three splits, in which there are 248,349 examples for training, 121,512 for validation, and 244,302 for testing. VQA2.0 is developed based on VQA1.0, and has about 1,105,904 image-question pairs, of which 443,757 examples are for training, 214,254 for validation, and 447,793 for testing. As a different setting, VQA2.0 provides annotations of similar image pairs as we mentioned above. We also collects image pairs in VQA1.0. For each example, we select its negative example that has the same question but different answers from both VQA1.0 and VQA2.0 datasets[3]. Then, we have about 197 thousand pairs for training. Since each question in these two datasets is associated with a list of answers, we use the most frequent answer as the anchor.

Model Configuration. In terms of the baseline model, we use the *Glove Embedding* (Pennington, Socher, and Manning 2014) as the word input with a dimension of 300. The dimension of the LSTM module is 2048, while the k and o in MFB fusion (Yu et al. 2017) are set to 5 and 1000, respectively. The dimensions of the last forward layer and the projections are set to 2048 and 300. The two hyper-parameters, α and β, are set to 0.25 and 0.01 after tuning. The initial

[3]We first select negative samples with exactly the same content, and then choose the negative samples with the most different answer lists.

Table 1: Ablation Studies. Trained on VQA2.0 *train* split and tested on the *val* split. "*FRCNN*" denotes the use of Faster R-CNN features.

Method	All	Num.	Yes/No	Others
Baseline	60.37	39.19	78.88	51.91
Baseline_EMB	60.74	39.51	79.11	52.40
Baseline_PIL	61.56	40.10	79.64	53.51
Baseline_PIL_HEM	62.01	41.22	80.03	54.21
Baseline+FRCNN	63.72	42.72	81.96	55.47
Baseline_EMB+FRCNN	64.00	43.42	81.69	56.00
Baseline_PIL+FRCNN	64.58	44.19	82.35	56.46
Baseline_PIL_HEM+FRCNN	64.63	44.27	82.29	56.59
PIL Setting	All	Num.	Yes/No	Others
$\alpha = 0.01,\ \beta = 0.25$	61.56	40.10	79.64	53.51
$\alpha = 0.01,\ \beta = 1$	61.06	39.64	79.14	53.14
$\alpha = 0.01,\ \beta = 0.5$	61.31	39.70	79.21	53.41
$\alpha = 0.01,\ \beta = 0.1$	61.48	40.25	79.52	53.38
$\alpha = 0.01,\ \beta = 0.05$	60.84	39.41	79.10	52.38
$\alpha = 0.25,\ \beta = 0.25$	61.50	40.37	79.51	53.36
$\alpha = 0.001,\ \beta = 0.25$	61.46	40.07	79.46	53.47

learning rate is 7e-4, which is halved after every 25,000 steps. The batch size is 64 and the maximum training step is set to 150,000. The optimizer we used is *Adam* (Kingma and Ba 2014). For the implemented the state-of-the-arts, we follow their default settings. During experiments, we use two types of visual inputs, *i.e.*, the last feature map of ResNet-152 (He et al. 2016) with a size of $14 \times 14 \times 2048$ and the regional features released by (Anderson et al. 2018) with a size of 36×2048. For simplicity, we denote them as *CNN* and *FRCNN*, respectively.

Experiment Analysis

Ablation Studies. We first examine different designs proposed in this paper, the results of which are shown in Tab.1. Here, *EMB* means that only the answer embedding loss is used as the regularization. *PIL* denotes the complete use of Pairwise Inconformity Learning, and *HEM* is the hard example mining. From Tab.1 we can see that, all designs are beneficial for the model performance, and considering that it requires very few extra parameters, these improvements are valuable. Meanwhile, with the number of training example increases, the improvement by PIL will be more obvious, as shown in Tab.5. Tab.1 also shows the tuning results of the two hyper-parameters α and β for the answer embedding loss and the triplet loss, as defined in Eq.8. The values in the first column are the setting we used for other experiments. With this setting, the embedding loss will give to more gradients to the model than that of the triplet loss. Our understanding is that the embedding loss not only carries the functionality of maintaining the multi-modal space, but also affects the projections of joint features. So, the value of its hyper parameter should be larger. Another observation is that if the values of hyper parameters are too large, it will impair the overall performance. One possible reason is that the question in VQA1.0 and VQA2.0 often has multiple answers, and a too strong regularization might reduce the diversity of the model prediction, making it hard to reach the optimal performance.

Table 2: Comparisons with other negative-learning scheme. Trained on VQA2.0 train set and tested on the validation set.

Method	All	Num.	Yes/No	Others
Baseline	60.37	39.19	78.88	51.91
Baseline+***l2_Reg.*** (α=1e-3)	59.47	37.64	77.93	51.22
Baseline+ ***Contra. Loss***(α=1e-3)	59.58	38.98	77.58	51.42
Baseline+***kl-D Reg.*** (α=5e-3)	59.18	38.20	77.29	50.94
Baseline+f_v-*l2_Reg.**(α=1e-4)	59.41	37.74	77.75	51.21
Baseline+f_v-PIL*	61.32	39.82	79.62	53.11
Baseline_PIL	61.56	40.10	79.64	53.51

*The regularization is implemented on the attention feature f_v.

Table 3: Applications on the State-of-the-Arts. Trained on VQA2.0 *train+val* split, and tested on *test-dev* split.

Method	All	Num.	Yes/No	Others
CNN+LSTM	54.22	-	-	-
CNN+LSTM-PIL	56.62	36.18	78.46	49.83
SAN	58.93	37.52	78.34	50.12
SAN-PIL	61.42	39.67	77.62	52.55
SAN_FRCNN	62.14	41.75	78.91	52.45
SAN_FRCNN+PIL	65.07	43.20	81.15	56.34
BUA_FRCNN	65.32	44.21	81.82	56.05
BUA_FRCNN+PIL	66.23	45.20	82.63	57.23

Comparisons with Other NL Schemes. We further compare the proposed PIL with other negative learning schemes in Tab.2. Here, "*l2_Reg.*" means that an $l2$-distance regularization is directly used to increase the semantic distances between the joint features f_j of pos./neg. examples, while "f_v-*l2_Reg.*" denotes that the regularization is implemented on the attention visual feature f_v. "*kl-D Reg.*" refers to the *KL-divergence*, which is used to increase the distribution divergence between predictions of positive and negative examples. "*Contra. Loss*" denotes the contrastive loss proposed in (Hadsell, Chopra, and LeCun 2006), which is also deployed on the joint features. "f_v-PIL" denotes the implementation of the proposed PIL on the attention features, and its setting is the same with that of PIL. From Tab.2, we find that, by directly regularizing the joint features, these negative learning schemes can have a counterproductive effect on the model performance, which confirms the hypothesis we made in the method section. The result of "f_v-PIL"also suggests that simply increasing the distances among visual features is insufficient for improving the role of visual features.

Plugging into State-of-the-arts. To verify the generalization of the proposed PIL, we deploy it on three state-of-the-arts, the results of which are shown in Tab.3. From this table, we find that the improvements of PIL on these models are

Table 4: Comparisons with the state-of-the-arts on VQA1.0. "$\mathcal{F}$" denotes the use of Faster R-CNN features, and "$\mathcal{G}$" means that the model uses *Visual Genome* (Krishna et al. 2016) as the additional training data.

VQA1.0 test-dev	All	Yes/No	Num.	Other
SAN (Yang et al. 2016)	58.7	79.3	36.6	46.1
DAMN (Andreas et al. 2016)	59.4	81.1	38.6	45.4
HiCo (Lu et al. 2016)	61.8	79.7	38.7	51.7
MCB (Fukui et al. 2016)	64.2	82.2	37.7	54.8
DAN$^{\mathcal{G}}$ (Nam, Ha, and Kim 2017)	64.3	83.0	39.1	53.9
MLB$^{\mathcal{G}}$ (Kim et al. 2017)	65.1	**84.1**	38.2	54.9
MFB (Yu et al. 2017)	**65.9**	84.0	**39.8**	**56.2**
Baseline (Ours)	64.7	83.2	37.9	54.9
Baseline_PIL (Ours)	66.1	84.5	39.5	56.4
Baseline$^{\mathcal{F}}$ (Ours)	66.9	**84.8**	40.8	57.5
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.6**	83.7	**41.4**	**59.7**
VQA1.0 test	All	Yes/No	Num.	Other
SAN (Yang et al. 2016)	58.9	-	-	-
DAMN (Andreas et al. 2016)	59.4	-	-	-
HiCo (Lu et al. 2016)	62.1	-	-	-
MCB$^{\mathcal{G}}$ (Fukui et al. 2016)	64.2	82.2	37.7	54.8
DAN$^{\mathcal{G}}$ (Nam, Ha, and Kim 2017)	64.3	83.0	39.1	53.9
MLB$^{\mathcal{G}}$ (Kim et al. 2017)	65.1	**84.1**	38.2	54.9
MFB$^{\mathcal{G}}$ (Yu et al. 2017)	**65.9**	84.0	**39.8**	**56.2**
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.7**	**83.7**	**41.3**	**59.6**

even higher than that of our baseline model. Such a result not only indicates the performance improvements brought by PIL, but also suggests that PIL can help models obtain more discriminative visual information. So its benefits will be more apparent when applied to less-than-good models. Notably, the additional parameters increased by PIL are only 5.4%, 2.4% and 1.8% for three models.

Comparing Baseline+PIL to State-of-the-arts. We further evaluate the baseline model with the proposed PIL on VQA1.0 and VQA2.0 datasets. The results are given in Tab.5. The first observation is that the proposed baseline is a simple yet powerful model. Without the use of additional datasets like Visual Genome, it has already reached the most advanced performance in VQA. Considering that any minor improvement is valuable on these two highly competitive datasets, PIL really takes effect. After deploying PIL, the baseline model outperforms *BUA*, the best state-of-the-art, by 2.23%, which is indeed significant in VQA. Notably, PIL only requires about 1.2% of parameters in addition.

Is the Impact of Knowledge Inertia Reduced?

We further examine whether the proposed PIL can strengthen the role of visual information in prediction, thereby reducing the impact of knowledge inertia. We first draw the prediction distributions on the 10 most common answers in Fig4. Here, "Ground Truth" denotes the default answer distribution of VQA2.0 val set. "Baseline" and "Baseline_PIL" are the prediction distributions by the model with and without PIL, respectively. Compared to the default distribution, these 10 answer categories receive more predictions from the model with a common VQA setting, which means that the model tends to choose the most frequent answers as their predictions. With PIL, the distribution of these ten answers is closer to the default distribution, which means that in more cases the model answers the question based more on the visual content and less effected by the language prior. So, we deduce that the influence of knowledge is alleviated.

Table 5: Comparisons with the state-of-the-arts on VQA2.0.

VQA2.0 test-dev	All	Yes/No	Num.	Other
MFB (Yu et al. 2017)	64.16	80.95	40.73	54.62
MF-SIG$^{\mathcal{G}}$ (Zhu et al. 2017)	64.73	81.29	42.99	55.55
BUA$^{\mathcal{G}}$ (Teney et al. 2018)	62.07	79.20	39.46	52.62
BUA$^{\mathcal{F},\mathcal{G}}$ (Teney et al. 2018)	**65.32**	**81.82**	**44.21**	**56.05**
Baseline (Ours)	63.53	80.40	41.16	54.23
Baseline_PIL (Ours)	65.07	81.15	43.20	56.34
Baseline $^{\mathcal{F}}$(Ours)	66.21	82.38	45.03	57.24
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.53**	**83.96**	**46.50**	**58.29**
VQA2.0 test	All	Yes/No	Num.	Other
LSTM-Q(Goyal et al. 2017)	54.22	73.46	35.18	41.83
MCB (Fukui et al. 2016)	62.27	78.82	38.28	53.36
MF-SIG$^{\mathcal{G}}$ (Zhu et al. 2017)	65.84	81.85	43.64	57.07
BUA$^{\mathcal{G}}$ (Teney et al. 2018)	62.27	79.32	39.77	52.59
BUA $^{\mathcal{F},\mathcal{G}}$ (Teney et al. 2018)	**65.67**	**82.20**	**43.90**	**56.26**
Baseline_PIL (Ours)	65.10	81.38	43.33	56.24
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.65**	**83.91**	**46.01**	**58.46**

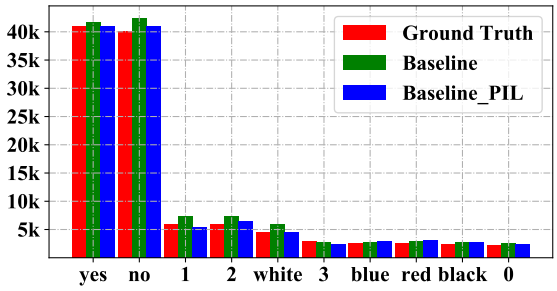
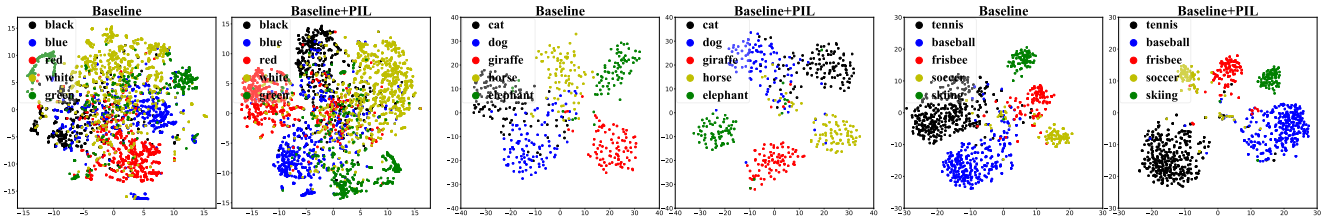


Figure 4: Distributions of the top-10 answers on the VQA2.0 *val* set.The *kl*-divergences between the ground truth and the baseline prediction on 3,000 answer categories is 0.06, and becomes smaller by adopting PIL (0.04).

Next, we visualize the joint features of baseline models with and without PIL by *t-SNE* (Maaten and Hinton 2008) in Fig.5. From Fig.5, it can be observed that PIL helps the model produce more discriminative joint features. Under the same category, the content of questions is relatively similar. So the larger differences among joint features indicate that the visual features are more discriminative for each answer category. We can therefore confirm that the role of visual information is enhanced in prediction.

To obtain a deeper insight into the model's prediction process, we visualize the attention maps with and without PIL in



(a) Answers of Colors (b) Answers of Animals (c) Answers of Sports

Figure 5: Visualizations of Joint Features without and with the proposed Pairwise Inconformity Learning by t-SNE.

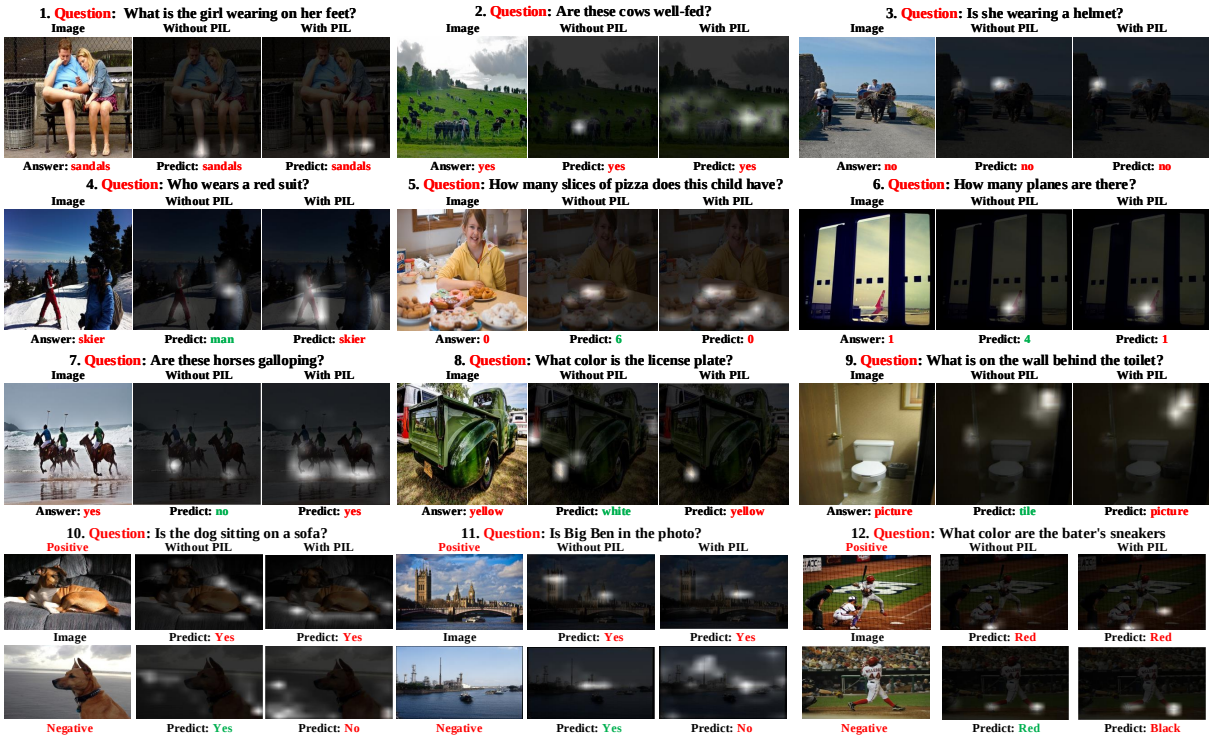


Figure 6: Visualizations of attentions with and without the proposed Pairwise Inconformity Learning (PIL). The red color in predictions refers to the correct answer while the green color denotes the incorrect ones.

Fig.6. Clearly, there are two typical cases of knowledge inertia in these examples. The first case is that, the model does not focus on the correct visual entities but can still predict the answers based on their "past experience", *e.g.* exp.(1)-(3). Although such a problem does not lead to a decline in the model performance, it still impairs the credibility of the model prediction. The second case is that, even if the model finds out the correct answer entity in the image, it will still make a wrong prediction. *e.g.* exp.(6)-(9). We also observe that under the example-pair setting, *e.g.* exp.(10)-(12), the baseline model often makes the same prediction for similar images. With PIL, such a case has been alleviated to a large extend, and the model can predict the answer more based on what they are currently "seeing". As shown in the rest examples, PIL also helps model produce attentions more accurately. In sum, PIL does strengthen the role of visual information and reduces the influence of knowledge inertia.

Conclusion

In this paper, we address the issue of knowledge inertia in Visual Question Answering, which is mainly caused by the strong language priors. To address this issue, we propose a Pairwise Inconformity Learning (PIL) to strengthen the importance of visual features by increasing the discrepancies between the example pairs. Its novelties includes the multimodal embedding learning, a dynamic-margin based triplet loss. PIL is also the first to use the image pair setting proposed in VQA2.0 to solve the strong language priors. Experimental results shows that with a negligible increase of parameters, our scheme can help the model to achieve a signif-

icant performance improvement. More importantly, statistic results and qualitative analyses prove that our scheme can reduce the impact of knowledge inertia.

Acknowledgments

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. *computer vision and pattern recognition*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, 6.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Goyal, Y.; Khot, T.; Summersstay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6325–6334.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *null*, 1735–1742. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Kim, J.; On, K. W.; Lim, W.; Kim, J.; Ha, J.; and Zhang, B. 2017. Hadamard product for low-rank bilinear pooling. *international conference on learning representations*.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

Liao, S. 2002. Problem solving and knowledge inertia. *Expert Systems With Applications* 22(1):21–31.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, 289–297.

Ma, L.; Lu, Z.; and Li, H. 2015. Learning to answer questions from image using convolutional neural network. *arXiv preprint arXiv:1506.00333*.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.

Nam, H.; Ha, J. W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Teney, D.; Anderson, P.; He, X.; and Den Hengel, A. V. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *computer vision and pattern recognition*.

Wang, P.; Wu, Q.; Shen, C.; and Hengel, A. v. d. 2016. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. *arXiv preprint arXiv:1612.05386*.

Wu, Q.; Wang, P.; Shen, C.; Dick, A.; and van den Hengel, A. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3.

Zhu, C.; Zhao, Y.; Huang, S.; Tu, K.; and Ma, Y. 2017. Structured attentions for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3.