Understanding Attention for Vision-and-Language Tasks

Anonymous ACL submission

Abstract

Attention mechanism has been used as an important component across Vision-and-Language(VL) tasks in order to bridge the semantic gap between visual and textual features. While attention has been widely used in VL tasks, it has not been examined the capability of different attention alignment calculation in bridging the semantic gap between visual and textual clues. In this research, we conduct a comprehensive analysis on understanding the role of attention alignment by looking into the attention score calculation methods and check how it actually represents the visual region's and textual token's significance for the global assessment. We also analyse the conditions which attention score calculation mechanism would be more (or less) interpretable, and which may impact the model performance on three different VL tasks, including visual question answering, text-to-image generation, text-and-image matching (both sentence and image retrieval). Our analysis is the first of its kind and provides useful insights of the importance of each attention alignment score calculation when applied at the training phase of VL tasks, commonly ignored in attention-based cross modal models, and/or pretrained models.

1 Introduction

The relative maturity and flexibility of deep learning allow us to build upon the success of computer vision and natural language processing to face many complex and multimodal Vision-and-Language (VL) tasks, such as Visual Question Answering (VQA), Text-and-Image Matching (T&I Match), or Text-to-Image Generation (T2I Gen). For these VL tasks, it is crucial to effectively align the multimodal information in both visual and linguistic domains. For example, to pick the right answer in VQA, the model should empower information from the input image, together with aligning the linguistic meanings with visual clues.

Attention mechanism (Bahdanau et al., 2015; Luong et al., 2015a) has been used as an important component across a wide range of VL models; from the early-stage attention-based fusion VL models (Xu et al., 2018a; Wang et al., 2019; Yang et al., 2016; Shih et al., 2016) to the recent VL multimodal transformer-based pretrained models (Lu et al., 2019; Li et al., 2020b; Hu et al., 2021). Those attention-based VL models mainly focus on 1) exploring new features to represent visual and linguistic information as an input of attention layer, 2) deciding the position or the number of attentions in the model, or 3) investigating the interpretability of attention distribution on VL tasks by emphasising the specific image regions or textual tokens.

While such approaches and investigations have resulted in interesting findings in different aspects of VL tasks, the attention alignment calculation between vision and language modalities has been less explored. However, the alignment calculation is directly linked to the main purpose of using attention mechanisms in VL tasks, which is to effectively bridge and align two different visual and linguistic informations. In other words, the essence of the attention mechanism in VL tasks is the alignment score calculation, as it quantifies the amount of "Attention" that the visual features would place on each of the language representations (or linguistic features would empower on the specific visual regions) when bridging the semantic gap between visual and language features. Most existing VL models directly apply the two attention alignment functions, a general and a dot-product (Luong et al., 2015a), which are commonly used in several NLP tasks. Since Vaswani et al. (2017) proposed a scaled dot-product for the transformer with full attention, almost every VL paper has directly applied those three attention alignment score functions. Instead, little work has been done towards understanding the role of attention alignment calculation methods applied to bridge visual and linguistic features, and exploring the impact on different VL model performance.

In order to address this limitation, the overarching goal of this research is to perform an extensive and systematic assessment of the effect of a range of attention alignment mechanisms pertaining to VL tasks, including three major VL tasks: Visual Question Answering (VQA), Text-and-Image Matching (T&I Match), and Text-to-Image Generation (T2I Gen). Towards that end, we systematically analyse the impact of the position of query and key in attention alignment on VL tasks. We investigate the following three questions: i) Which attention alignment score calculation yields the most benefit in VL tasks? ii) What if we linearly transform the query Q instead of the key K (or vice versa) before the multiplication? For example, assume the textual feature T is a query Q, and the image feature I is a key K. We analyse the impact of linear transforming Q or K in alignment score calculation. iii) Do the attention alignment calculation techniques with better performance provide better attention distribution interpretability?

In brief, our **main contributions** are as follows: **1)** We conduct a comprehensive analysis of the role of attention alignment score calculation in VL tasks (including three widely-used VL tasks, such as Visual Question Answering, Text-and-Image Matching, and Text-to-Image Generation). **2)** We perform a comparative analysis of the position of query and key (language and visual feature) for the alignment calculation. **3)** We evaluate the interpretability of the best and worst attention alignment calculation models. **4)** We make the code and the data publicly available to encourage reproducibility of results.

2 Related Works

Several NLP research efforts investigate the interpretability of attention mechanisms; Jain and Wallace (2019) assessed the interpretability by empirically inspecting the learned attention weights in NLP models, and Sun and Lu (2020) explored the internal mechanism of attention for text classification by analyzing gradient update process. Multimodal VL models directly adopted the attention mechanism to bridge the visual and linguistic modal information. In this section, we describe an overview of related workS on the role of attention mechanisms on different VL tasks, including Text-to-Image Generation, Text-and-Image Matching, Visual Question Answering, and Text-based Visual Question Answering.

Text-to-Image Generation AttnGAN (Xu et al., 2018a) first proposed word-level attention with iterative image generation. They used dot-product for measuring the alignment between visual subregions and word tokens, based on which the word-context vector for each image subregion was produced to guide its generation process. Many of the later approaches directly adapted the dot-product attention from AttnGAN while focusing on improving other components in the architecture (Zhu et al., 2019; Yin et al., 2019; Qiao et al., 2019b,a; Li et al., 2019, 2020a; Han et al., 2020; Pande et al., 2021). A few models apply an element-wise multiplication (Qiao et al., 2019a,b) or cosine similarity (Zhang et al., 2021a) as an attention alignment measurement. However, none of those models explore any alternatives by a comparative analysis.

Text-and-Image Retrieval Similar to the Text-to-Image generation, a cosine similarity based attention alignment was applied by SCAN (Lee et al., 2018), the most widely-used baseline model in Text-and-Image Retrieval. Since then, many studies directly adapted its attention alignment calculation with no or minor change (Liu et al., 2019; Chen and Luo, 2020; Chen et al., 2020; Diao et al., 2021; Dong et al., 2021). They applied text-to-image (t2i) and image-to-text(i2t) attention in two separate variants to filter the cross-modal relevant representations for later image-sentence matching. Some other approaches applied (scaled) dot-product instead (Wang et al., 2019; Liu et al., 2020; Wei et al., 2020; Fei et al., 2021). However, they also do not have proper justifications for selecting the alignment calculation method.

Visual Question Answering (VQA) Both textual query-guided image attention and image-guided textual query attention have been commonly used in VQA approaches, which utilised one modality to guide the focus on the other. Several categories of alignment calculations or their variants were included, such as adapting neural networks (Yang et al., 2016; Zhu et al., 2016; Patro and Namboodiri, 2018; Anderson et al., 2018) or applying (scaled) dot-product (Hudson and Manning, 2018; Yu et al., 2019; Gao et al., 2019; Huang et al., 2020; Rahman et al., 2021; Zhang et al., 2021b; Guo et al., 2021) etc. Despite lots of alignment selections, their effect in VQA remains less explored.

Text-based Visual Question Answering Recent TextVQA approaches directly augmented existing

VQA models, and their cross-modal attention with additional OCR inputs (Singh et al., 2019; Biten et al., 2019a,b; Wang et al., 2020). Both early-stage model *M4C* (Hu et al., 2020) and the most recent pretrained model *TAP* (Yang et al., 2021) fed the question, image and OCR text together into a multimodal transformer and jointly encoded them via scaled-dot product attention in the transformer encoder. As with other VL tasks though, there is a lack of research on exploring the most effective cross-modal attention alignment for this task.

Hence, we examine the impact of attention alignment score calculation in different VL downstream tasks, including Visual Question Answering, Text-and-Image Matching, and Text-to-Image Generation. We select the most widely adopted baseline models for each downstream task; AttnGAN for the Text-to-Image Generation, SCAN for the Text-to-Image Matching, MAC for the Visual Question Answering, and M4C for the Text-based Visual Question Answering, and evaluate the impact of attention alignment score measurement.

3 Attention Alignment Mechanism

There are various attention mechanisms applied in different multimodal VL downstream tasks. Two commonly used approaches are the cross attention and the self-attention. First, the cross attention is performed between visual and textual inputs. More specifically, given a sequence of textual features $T = \{t_1, t_2, t_3, \ldots, t_M\}$ and image features $I = \{i_1, i_2, i_3, \ldots, i_N\}$, it takes T as the query Q and I as the key K (or vice versa) to compute attention and context vectors c as the attended representations of the input elements in the following way:

$$a_{xy} = f(Q_x, K_y) \tag{1}$$

$$\alpha_{xy} = \frac{exp(a_{xy})}{\sum_{y=1}^{n_K} exp(a_{xy})} \tag{2}$$

$$c_x^K = \sum_{y=1}^{n_K} \alpha_{xy} K_y \tag{3}$$

where f is a function to calculate attention score, n_K is the number of elements in K, and c_x^K is the context vector of K with respect to the x-th element of Q. The second approach, self attention (Vaswani et al., 2017), is performed over all inputs from both modalities. In other words, the approach combines T and I as a complete sequence $S = T \cup I$, and converts all elements in S into Q, K and V via learnable matrices, which are used to compute attention by multiple heads in the following way:

$$\text{Attention}(Q, K, V) = \text{Softmax}(f(Q, K))V \tag{4}$$

where the results from different heads are combined together. Then, it applies layer normalization, residual connections and fully connected layers in order to obtain the attended representation of the input tokens. With both approaches, we explore the effect of the attention alignment calculation f for different VL tasks with the following five different alignment score functions. Note that we also include **Cosine similarity**-based attention for only Text-and-Image Matching as it is widely used in that specific domain.

Dot product attention It was proposed in NMT (Bahdanau et al., 2015) to compute vector similarity between encoder hidden states and decoder hidden states. This function (Luong et al., 2015b) has been widely adopted as f in the cross attention mechanism as shown in Equation 1.

$$f(Q, K) = QK \tag{5}$$

Scaled dot product attention The higher dimension of data representation would lead to the smaller gradient of softmax function. Hence, the scaling factor was introduced by Vaswani et al. (2017), and applied to the self attention-based VL approaches as represented in Equation 4.

$$f(Q, K) = \frac{QK}{\sqrt{d}} \tag{6}$$

General attention Along with dot product attention, general attention (Luong et al., 2015b) received lots of interest as an alternative alignment calculation method that computes attention score using an extra learnable matrix to linearly transform K into the same embedding space as Q. This can be considered as one of the neural network based methods mentioned in Section 2.

$$f(Q, K) = QWK \tag{7}$$

There are several variants of neural network based general attention calculation methods. First, **Biased general attention** is introduced by Sordoni et al. (2016) using more bias towards more important keys regardless of the query context.

$$f(Q, K) = Q(WK + b) \tag{8}$$

Secondly, **Activated general attention.** Ma et al. (2017) applies an additional nonlinear activation term, which is able to amplify the emphasis on query elements that are highly relevant to the key.

$$f(Q, K) = act(Q(WK + b)) \tag{9}$$

In this paper act is the ReLU activation since it is a widely used function in VL downstream tasks.

4 Vision-Language Models

We use publicly available implementations of the most widely adopted VL baseline models in order to train and evaluate different attention alignment score calculation for three different VL tasks: (i) **AttnGan** for Text-to-Image Generation (T2I Gen), (ii) **SCAN** for Text-and-Image Matching (T&I Match), (iii) **MAC** and **M4C** for each Visual Question Answering (VQA) and Text-based Visual Question Answering (TVQA).

4.1 T2I Gen: AttnGAN

The goal of text-to-image generation is to generate a visually realistic image that matches a given text description. The AttnGAN (Xu et al., 2018b) generates images by using multiple generators with the attention mechanisms. To improve the image quality at each step, a cross attention mechanism is performed between caption words and image regions, and it produces the attended word context for each image region. Given a caption of M words, an image with N sub-regions would be generated by an upsampling network. The words and image regions are represented as d-dimensional vectors $\{t_m\} \in T$ and $\{i_n\} \in I$ respectively. Then image representation I is applied as Q the query and caption representation T is applied as K the key for the cross attention mechanism (Equations 1, 2, 3), where the dot product attention score calculation is used as f. The resultant textual context would be fused with word region representations as a guide for the generator at the next time step to focus on different words. Note that we evaluate different alignment calculation methods as f to investigate the impact of the image generation performance. We fix I as Q and T as K, and replace the dot product with other alignment score calculations.

4.2 T&I Match: SCAN

Text-and-image matching (a.k.a. Text-and-image retrieval) refers to measuring the visual-semantic similarity between a sentence and an image. The SCAN model (Lee et al., 2018) performs a pairwise cross attention between image regions and caption words for fine-grained T&I Match. This can be done in two directions. Given a caption of M words and an image having N detected objects, d-dimensional representations $\{t_m\}$ and $\{i_n\}$ are obtained as T and I respectively. To obtain the attended image context for each caption word, the cross attention mechanism (described in Equations 1, 2) is applied with T being the query Q and I being the key K, and an alignment score is measured by using cosine similarity between each caption word and its image context. These alignment scores would be aggregated via a pooling function as the final alignment score between the given image and caption. Such scores can be obtained by using T as K and I as Q to calculate the sentence context for each image region. In experiments, we fix T as Q and I as K, and replace the cosine similarity with other alignment score calculations.

4.3 VQA

We explore two VQA downstream tasks, Visual Question Answering with compositional reasoning and Text-based Visual Question Answering.

4.3.1 VQA: MAC

First, we focus on the visual question answering task that requires responding to natural language questions about images, specifically with a compositional and structured nature. The MAC circuit (Hudson and Manning, 2018) applies a cross attention mechanism to answer a question based on a given image. Instead of computing attention between textual and visual input, MAC introduces a d-dimensional learnable control state e as a guidance for MAC cells to selectively attend to different aspects of inputs at each time step. Within each MAC cell, there is a control unit to attend to the question words and a read unit to attend to the image regions. Given a question of M words and an image having N detected objects, d-dimensional representations $\{t_m\}$ and $\{i_n\}$ are obtained as T and I respectively. Instead of using Equation 1, the control unit applies e as Q and T as K to compute the attention score in the following way:

$$a_y = W'(f(Q, K_y)) + b' \tag{10}$$

where f indicates element-wise dot product multiplication to obtain a d-dimensional similarity vector, and W' and b' are learnable parameters to output a scalar as the score. Then the control unit follows Equations 2, 3 to obtain textual context as an update for e. Similar to the control unit, the read unit applies e as Q and I as K to obtain the question-guided visual context from the image, which is later aggregated to predict an answer. Therefore the read unit can be considered as a main component in MAC that involves multimodal alignment. Hence, for the evaluation, we fix the control state e (which majorly contains textual question information) as Q and image-based knowledge graph I as K, and adapt the focused attention alignment calculation methods f with the element-wise multiplication manner in the read unit.

4.3.2 TVQA: M4C

Secondly, Text-based visual question answering (TVQA) is an extension of VQA, which requires the model to read text over the image to answer the questions. The M4C model (Hu et al., 2020) applies a multimodal transformer over all input modalities to perform iterative answer prediction for the TextVQA task. More specifically, given a question of M words, an image having N detected objects and O detected OCR tokens, d-dimensional representations $\{s_m^{ques}\}$, $\{s_n^{obj}\}$ and $\{s_o^{ocr}\}$ are obtained as input sequence S. The self-attention mechanism (Equation 4) with scaled dot product attention is applied over S, the sequence of all $M + N + O$ entities. In this way, both intra-modal interactions and inter-modal interactions are captured to aggregate the input to form an answer prediction via classical transformer layers. Similarly to other tasks, we replace the scaled dot product attention calculation with the other aforementioned options for f to investigate the impact in TVQA.

5 Evaluation Setup[1]

5.1 Datasets

We conducted experiments on three VL task datasets. The dataset statisitcs can be found in Appendix Table 4. We followed the work of the base models, including AttnGAN (Xu et al., 2018b), SCAN (Lee et al., 2018), MAC (Hudson and Manning, 2018), M4C (Hu et al., 2020) for dataset preprocessing and dividing for train/dev/test sets.

5.1.1 T2I Gen

Two benchmark datasets are used: **Caltech-UCSD Birds 200 (CUB)**[2] and **MS-COCO**[3]. CUB has 11,788 images of 200 bird categories downloaded from the Flickr website, each with 10 textual captions. MS-COCO provides 123,287 images of complex everyday scenes with 5 manually written textual descriptions per image. We use a train/test split of 8,855/2,933 and 82,783/15,000 images respectively for CUB and MS-COCO.

5.1.2 T&I Match

Flickr30k[4] contains around 31k images collected from the Flickr website with 5 crowd-sourced captions per image. We test on Flickr30k with train/dev/test split of 29k/1k/1k images and on MS-COCO (as described above) with 29k/1k/1k images.

5.1.3 VQA

We have two VQA tasks: 1) Visual Question Answering with compositional reasoning, and 2) Text-based Visual Question Answering. We used **CLEVR**[5] and **TextVQA**[6] respectively. CLEVR contains 100,000 synthetic images of 3D shapes with 999,968 questions/answers in total. We use a subset of 70,000 images with 699,989 QAs for training, 15,000 images with 149,991 QAs for validation and 15,000 images with 149,988 QAs for test. TextVQA consists of 45,336 questions asked by (sighted) humans on 28,408 images from the Open Images dataset (Krasin et al., 2017). We use the original split: 21,953 images with 35,602 QAs, 3,166 images with 5,000 QAs and 3,289 images with 5,734 QAs for training, validation and test.

5.2 Evaluation Metrics

We describe evaluation metrics used for assessing the impact of attention alignment mechanism for each VL task.

5.2.1 T&I Match: R@K

We measure the performance of sentence retrieval and image retrieval by recall at K (R@K), which is defined as the percentage of queries that get the correct item at the closest K points to the query. The higher the value, the better the performance.

5.2.2 T2I Gen: Inception Score(IS) & FID

The evaluation measurement we use is **Inception Score (IS)** which seeks to capture the image quality and image diversity properties of a collection of generated images. The higher the inception score, the better the model. **Frechét Inception Distance (FID)** measures the similarity between the generated images and the real images by comparing their

[1] The implementation details of our experiments can be found in Appendix A.1

[2] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[3] https://cocodataset.org/#home

[4] http://shannon.cs.illinois.edu/DenotationGraph/

[5] https://cs.stanford.edu/people/jcjohns/clevr/

[6] https://textvqa.org/dataset

Attention	Sentence Retrieval						Image Retrieval					
	Flickr30K			MS-COCO			Flickr30K			MS-COCO		
	R@1	R@10	Rsum	R@1	R@10	Rsum	R@1	R@10	Rsum	R@1	R@10	Rsum
cosine similarity$^\diamond$	62.4	93.3	243.8	**61.4**	94.5	243	43.9	81.8	199.9	45.7	88.2	212.5
dot product	62.1	92.1	240.4	59.7	**95.2**	243.5	44.8	**82.1**	200.6	**46.0**	87.9	212.6
scaled dot product	63.0	**93.8**	244.9	59.3	95.1	**243.7**	44.9	81.9	200.4	45.8	**88.3**	**213.4**
general*	**63.2**	93.6	**245.0**	59.8	**95.2**	**243.7**	**46.7**	81.8	**201.8**	45.6	87.8	212.1
general†	56.6	90.1	229.9	53.8	93.1	231.5	38.4	77.0	182.0	39.3	83.4	195.3
biased general*	56.6	89.8	230.3	52.2	91.7	227.0	39.6	77.3	185.0	38.7	83.4	194.5
biased general†	55.8	89.7	228.3	52.6	93.2	231.1	39.3	77.4	184.6	39.8	84.2	197.3
activated general	56.2	90.5	229.2	53.9	92.9	231.3	39.2	77.4	184.7	39.5	84.0	195.6

Table 1: R@1, R@10 and the sum of (R@1+R@5+R@10) on Flickr30K and MS-COCO for T&I Match. The definition of ◇, *, † can be found in footnote 7. Q refers to caption words and K refers to image regions.

Attention	Acc.
dot product$^\diamond$	0.966
scaled dot product	**0.973**
general*	0.967
general†	0.962
biased general*	0.959
biased general†	0.963
activated general	0.971

(a) VQA on CLEVR

Attention	Acc.	ANLS
dot product	0.407	0.545
scaled dot product$^\diamond$	**0.419**	**0.554**
general*	0.407	0.546
general†	0.416	**0.554**
biased general*	0.412	0.553
biased general†	0.414	0.551
activated general	0.413	0.548

(b) TVQA on Text-VQA

Table 2: Results for VQA/TVQA. The definitions of ◇, *, † are in footnote 7. For VQA, Q refers to the control state and K refers to image-based knowledge graph in read unit. For TVQA, Q and K are transformed union of all caption words, image object and OCR features.

Frechét distance between the maximum entropy distribution. Lower FID indicates higher similarity.

5.2.3 VQA

For the **VQA with compositional reasoning**, overall accuracy is used to measure the performance of the VQA models. The higher the accuracy, the better the performance of the model. For the **TVQA**, it is designed for the VQA context where 10 ground truth answers are provided for each question-image pair. The accuracy of a single prediction is a soft score obtained by a vote of the 10 ground truth answers. Overall accuracy is obtained by taking the average across all instances. In addition, we use the Average Normalized Levenshtein Similarity (ANLS) score (Biten et al., 2019b), which aims to eliminate the dropped performance caused by OCR recognition error by comparing the string similarity between the ground truth answers and the prediction results.

6 Results

We analyse the impact of attention alignment mechanisms in different VL tasks, and explore the interpretability based on attention distribution.

6.1 Test Performance

A primary goal of this work is to identify the most effective and successful attention alignment calculation functions for VL tasks. Tables 1, 2, and 3 [7] detail the results of our experiments comparing performance of individual alignment functions with each VL models. Note that each table visualises the trends with a heatmap. The darker the colour of the cells, the better the performance.

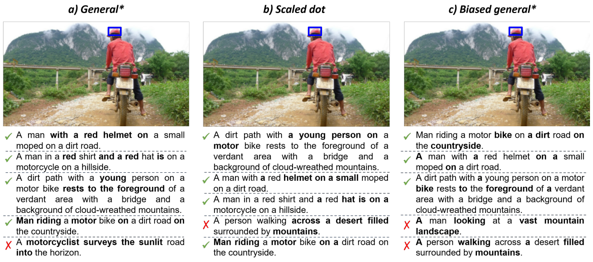
Observing the performance of the individual alignment functions on the T&I Match task in Table 1, we note that the original calculation function, cosine similarity, achieved quite good performance. However, scaled dot product and general* appears to be consistently the most effective function in both two Flickr30K and MS-COCO, indicating its effectiveness in both sentence retrieval and image retrieval. On the other hand, biased and activated general attention produce very low results in both R@K and Rsum values.

Table 2 details the performance of alignment functions in VQA. Note that we have two types of VQA, including VQA with compositional reasoning (CLEVR) and TVQA (Text-VQA). Surprisingly, both VQA and TVQA models produce the best performance with a scaled dot product alignment, highlighting the benefit in the visual question answering domain. We note that the activated general attention (ReLU activation) with the VQA performed well in the answer prediction, whereas the same function with the TVQA model produced one of the lowest ANLS scores. The general attention alignment function also provides opposite performance trends in two tasks; the detailed analysis for the position of key and query will be discussed in

[7] ◇ indicates the original attention alignment function used by the base models. * indicates $f(K, Q)$ (swapping query and key), and † indicates $f(Q, K)$ (without swapping query and key) for Equations 7 and 8

Attention	CUB		MS-COCO	
	IS	FID	IS	FID
dot product$^\diamond$	4.32	<u>**25.72**</u>	23.28	<u>**40.19**</u>
scaled dot product	4.31	25.74	23.84	42.33
general*	4.36	28.21	24.28	40.82
general†	4.26	26.94	24.63	42.45
biased general*	4.13	26.97	23.05	43.10
biased general†	4.30	25.89	<u>**25.24**</u>	43.64
activated general	<u>**4.41**</u>	28.39	23.56	42.65

Table 3: Results on CUB and MS-COCO for T2I Gen. The definitions of $\diamond$, *, $\dagger$ are in footnote 7. Q refers to caption words and K refers to image subregions.



a) General*

✓ A man **with a red helmet on** a small moped on a dirt road.

✓ A man in a **red** shirt **and a red** hat **is** on a motorcycle on a hillside.

✓ A dirt path with a **young** person on a motor bike **rests to the foreground** of a verdant area with a bridge and a background of cloud-wreathed mountains.

✓ **Man riding** a **motor** bike **on** a dirt road **on** the countryside.

✗ A **motorcyclist surveys the sunlit** road **into** the horizon.

b) Scaled dot

✓ A dirt path with **a young person on** a **motor** bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains.

✓ A man with **a** red **helmet on a small** moped on a dirt road.

✓ A man in a red shirt and **a** red **hat is on a** motorcycle on a hillside.

✗ A person walking **across a desert filled** surrounded by **mountains**.

✓ **Man riding** a **motor** bike **on a** dirt road on the countryside.

c) Biased general*

✓ Man riding a motor **bike** on **a dirt** road **on** the **countryside**.

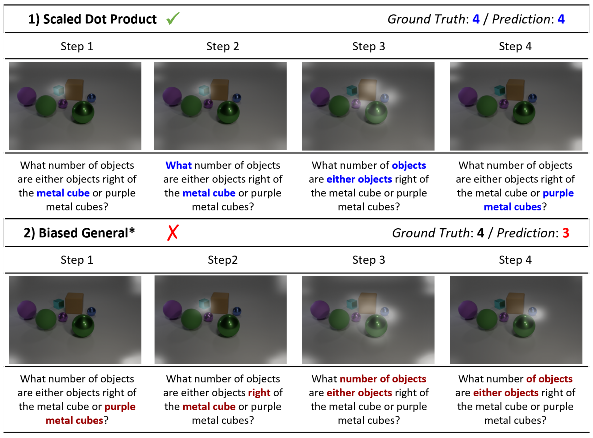
✓ **A** man with **a** red helmet **on a** small moped **on** a dirt road.

✓ A dirt path with **a** young person on a motor **bike** rests **to** the **foreground** of **a** verdant area with a bridge and a background of cloud-wreathed mountains.

✗ **A** man **looking** at a **vast mountain landscape**.

✗ **A** person **walking** across **a** desert **filled** surrounded by **mountains**.

Figure 1: Qualitative examples of T&I Match-MSCOCO with SCAN by different attention functions.



1) Scaled Dot Product ✓ — Ground Truth: 4 / Prediction: 4

Step 1: What number of objects are either objects right of the **metal cube** or purple metal cubes?

Step 2: **What** number of objects are either objects right of the **metal cube** or purple metal cubes?

Step 3: What number of **objects** are **either objects** right of the metal cube or purple metal cubes?

Step 4: What number of objects are either objects right of the metal cube or **purple metal cubes**?

2) Biased General* ✗ — Ground Truth: 4 / Prediction: 3

Step 1: What number of objects are either objects right of the metal cube or **purple metal cubes**?

Step2: What number of objects are either objects **right** of the **metal cube** or purple metal cubes?

Step 3: What **number of objects** are **either objects** right of the metal cube or purple metal cubes?

Step 4: What number **of objects** are **either objects** right of the metal cube or purple metal cubes?

Figure 2: Qualitative examples of VQA-CLEVR from the MAC trained by different attention functions.

Section 6.2. Considering that the nature of general VQA is different from that of TVQA, which mainly focus on OCR text input, it is unsurprising that the impact of alignment mechanism is opposite. Hence, it is remarkable to find that the scaled dot product achieved the best in both domains.

The result of T2I Gen presented in Table 3 produces quite different trends compared to other two tasks, including T&I Match and VQA. First, there is no alignment function that produces a consistent better performance in both evaluation metrics, IS or FID. While neural network-based alignment functions (i.e. general, biased and activated general) achieved higher IS scores than others, the dot product produced better FID scores than others. This trend can be seen in both CUB and MS-COCO. The scaled dot product obtains comparably good FID results, not in IS. The reason can be easily found if we understand the nature of IS and FID. IS focuses on the diversity of image, whereas FID measures the similarity between the ground-truth images and the generated images based on the textual description. Hence, if the alignment function between visual and textual information successfully works, it produces a better FID score. However, the better alignment function does not necessarily produce diverse objects in the image.

Based on all three downstream tasks, we can find the scaled dot product can be the best alignment calculation function that can successfully bridge the visual and textual information and can effectively work in both cross attention and self attention model, as it produces considerably and consistently better results across all six VL datasets.

6.2 Impact of Key and Query

We also investigated the impact of position of query and key in the attention alignment calculation process, especially when extra learnable weights and biases are involved. We explore the difference between linearly transforming the key K to multiply with the query Q ($f(K, Q) = KWQ$) and transforming the query Q to multiply with the key K ($f(Q, K) = QWK$) in general attention and biased general attention calculation. Specifically, we initially fixed the textual information as a query Q and visual information as a key K (stated in Equation 7 and 8) and swapped the position in different general attention alignment score measurements. In Table 1, 2, and 3, * indicates the functions with $f(K, Q)$, whereas $\dagger$ refers to those with $f(Q, K)$.

Table 1 shows that Flickr30K performed better with general* or biased general*, whereas MS-COCO does not have obvious trends. Similar patterns can be found in both cross attention mechanisms (VQA models, T2I Gen models), and self attention mechanism TVQA models. Interestingly but unsurprisingly, we note that there is no obvious and consistent performance improvement pattern in different positions of textual information (query Q) and visual information (key K) when it calculates the alignment. It depends on the specific downstream tasks and dataset. We can conclude that the way of calculating alignment is the crucial point in VL tasks, compared to the position/order of different modal information.

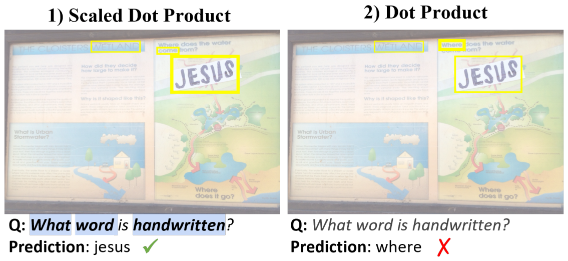


Figure 3: Qualitative examples of Text-VQA from the M4C trained by different attention alignment functions. Question words that receive attention weights greater than 0.01 are indicated in bold and coloured in blue.

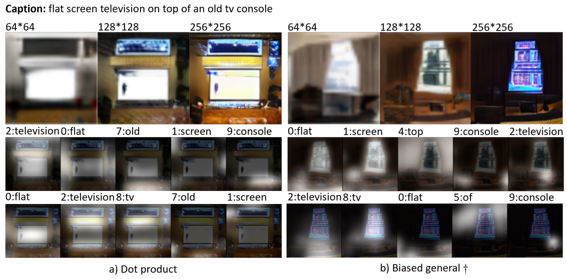


Figure 4: Qualitative examples of T2I Gen-MSCOCO with AttnGAN by different attention functions.

6.3 Qualitative analysis

We also evaluated and visualised the prediction interpretability of the best and worst attention alignment calculation functions for different VL tasks.

Figure 1 shows some examples for T&I Match. With general attention* and scaled dot product, the SCAN model is able to include 4 correct captions among the top 5 retrieved captions, while biased general attention can only extract 3 correct captions. With respect to the helmet indicated by the blue box in the image, both general attention and scaled dot product can put more focus on the words *red*/*helmet*/*hat* from the caption, whereas the biased general attention would rather focus on prepositions, determinants or other objects.

For VQA, Figure 2 shows a picture featuring several cubes and spheres. With a scaled dot product for attention score calculation, when the model focuses on the keyword *metal cubes* from the textual question, the only metal cube in the image is emphasized during the first two steps. Then, it correctly detects 4 objects from the image by highlighting the keywords *objects* and *either objects*. Additionally, the model looks for the *purple metal cube* from the picture as asked by the question, but it does not exist in the picture so none of the objects are highlighted at step 4. However, the model with the biased general attention tries to count *the number of objects* on the right of the metal cube in step 3 but it inaccurately focuses on the metal cube itself in addition to the correct ones. In the last step the model puts more focus on the farthest right objects, resulting in a wrong prediction of *3*. The qualitative analysis of TVQA is in Figure 3. The model with scaled dot product focused on the keywords *what*, *word* and *handwritten* to focus on the handwritten word *jesus* in the image and retrieved the correct OCR token with highest attention weight. However, with dot product attention, all the question words received little attention by the model (< 0.01), failing to find the appropriate OCR token in the image.

Figure 4 shows the images generated by AttnGAN using both the best and the worst attention, dot product and biased general† respectively. AttnGAN with a dot product, can generate a relatively more realistic image. From a low resolution picture, the model focuses on the words based on the following order, *television*, *flat*, *old*, *screen*, *console*, in order to refine the image to include the objects and corresponding features gradually. Compared to that, the biased general attention model generates a surrealistic image by focusing on *flat*, *screen*, *top*, *console*, *television* in the first step.

Overall, based on the qualitative analysis in different VL tasks, we reveal that the better attention alignment calculation function can produce better interpretability in terms of the prediction.

7 Conclusion

We systematically examined the role of attention alignment score calculation in vision-and-language tasks, including visual question answering, text-and-image matching, and text-to-image generation. We found that the scaled dot product function can be the best attention alignment calculation for either cross or self-attention in overall VL tasks while the appropriate position of visual and textual information may vary from different VL tasks/-datasets. In conclusion, we note that better vision-and-language information alignment leads to better task performance and interpretability. It is hoped that our analysis can provide a great insight for selection of the most effective attention alignment calculation for different VL benchmark tasks. We leave the systematical exploration of visual-textual attention interpretability to our future work.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019a. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019b. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4291–4301.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663.

Tianlang Chen and Jiebo Luo. 2020. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10583–10590.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1218–1226.

Xinfeng Dong, Huaxiang Zhang, Xiao Dong, and Xu Lu. 2021. Iterative graph attention memory network for cross-modal retrieval. *Knowledge-Based Systems*, 226:107138.

Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6639–6648.

Wenya Guo, Ying Zhang, Jufeng Yang, and Xiaojie Yuan. 2021. Re-attention for visual question answering. *IEEE Transactions on Image Processing*, 30:6730–6743.

Caren Han, Siqu Long, Siwen Luo, Kunze Wang, and Josiah Poon. 2020. Victr: Visual information captured text representation for text-to-vision multimodal tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3107–3117.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. Vivo: Visual vocabulary pre-training for novel object captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1575–1583.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7166–7176.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020a. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020b. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4068–4074. AAAI Press.

Sharad Pande, Srishti Chouhan, Ritesh Sonavane, Rahee Walambe, George Ghinea, and Ketan Kotecha. 2021. Development and deployment of a generative model-based framework for text to photorealistic image generation. *Neurocomputing*.

Badri Patro and Vinay P Namboodiri. 2018. Differential attention for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7680–7688.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019a. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in Neural Information Processing Systems*, 32:887–897.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019b. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514.

Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. 2021. An improved attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1653–1662.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Alessandro Sordoni, Philip Bachman, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *ArXiv*, abs/1606.02245.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018a. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Tao Xu, Pengchuan Zhang, Han Zhang Qiuyuan Huang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018b. Attngan: Fine-grained text to image generation with attentional generative adversarial networks.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pretraining for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8761.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2327–2336.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021a. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842.

Weifeng Zhang, Jing Yu, Wenhong Zhao, and Chuan Ran. 2021b. Dmrfnet: Deep multimodal reasoning and fusion for visual question answering and explanation generation. *Information Fusion*, 72:70–79.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Appendix

A.1 Experimental Settings

For **T&I Match: SCAN** (t-i) AVG models, all settings of hyper-parameters follow the configuration of the SCAN (Lee et al., 2018). The batch size is 128, the margin of triplet loss α is 0.2 and the threshold of maximum gradient norm for gradient clipping is 2. For Flickr30k models, the learning rate is set as 0.0002 for the first 15 epochs and then lowered to 0.00002 for another 15 epochs. Total training epochs are 30 and the best model is selected with the highest sum of R@K score. For MS-COCO models, we trained with a learning rate of 0.0005 for 10 epochs and then lowered the learning rate to 0.00005. The best model is selected with the highest sum of R@K score. Training epochs are 20. For **T2I Gen: AttnGan** model on CUB dataset, the batch size is set to be 20 and we trained with 400 epochs in total. On the MS-COCO dataset, the batch size is 14 and total epochs are 90. In addition to this, all settings are the same as the AttnGan (Xu et al., 2018b). For **VQA: MAC** models, the training epoch is set to be 8 and other hyperparameter settings are consistent with MAC(Hudson and Manning, 2018). More specifically, the batch size is 128, the learning rate is 0.0001 with 0.5 learning decay rate and the threshold of maximum gradient norm for gradient clipping is 8. For **TVQA: M4C** model on the Text-VQA dataset, we followed the exact same setting as M4C (Hu et al., 2020), applying the batch size of 128 and 100 epochs for training, All model variants would train to convergence within 80 epochs.

Tasks	**Dataset**	**Train**	**Dev**	**Test**
T2I Gen	CUB	8,855	-	2,933
	MS-COCO	82,783	-	15,000
T&I Match	Flickr30k*	29,000/145,000	1,000/5,000	1,000/5,000
	MS-COCO*	29,000/145,000	1,000/5,000	1,000/5,000
VQA	CLEVR*	70,000/699,989	15,000/149,991	15,000/149,988
	Text-VQA*	21,953/34,602	3,166/5,000	3,289/5,734

Table 4: Details of train/dev/test split for each dataset. Note that * indicates the dataset having different numbers for visual and textual inputs. It reports the number of images followed by the number of captions or question-answer pairs, separated by backslash (/).

In addition, we show the dataset split details in Table 4 and state the licenses of used assets in Table 5. All the artifacts we used either did not specify terms of use or limited the use to non-commercial research purpose. Since we conduct a systematic research study which is not for commercial purpose or application, our work is consistent with the terms of use of these assets.

Asset	License
Datasets	
CUB	CC BY 4.0
MS-COCO	CC BY 4.0
Flickr30k	Unknown
CLEVR	CC BY 4.0
Text-VQA	CC BY 4.0
Base Model Codes	
MAC	Apache License 2.0
SCAN	Apache License 2.0
M4C	BSD
AttnGAN	MIT

Table 5: Licenses of the assets used by the study.

A.2 Computing Infrastructure

All experiments for T2I Gen, T&I Match and VQA are conducted on a variety of cloud instances from Google Colab, with each utilising an NVIDIA Tesla T4 GPU of 16GB RAM. For TVQA the experiments are conducted utilising NVIDIA Titan RTX GPU with 24GB RAM, 16 Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz with 128GB RAM, and the operating system of Ubuntu 20.04.1.

A.3 Additional Qualitative Examples - VQA

We include more comparison examples for MAC model in this section to show the difference between scaled dot product and biased general* in the VQA context. In Figure 5, a question *what number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball* is raised towards an image with several cylinders, cubes and spheres. The MAC model with scaled dot product attention is able to correctly focus on the *brown matte object* from both the question and the image, while putting slight attention on *the brown thing on the right side* as mentioned in the question. Then in step 3 and 4 the model is able to locate the *small metallic thing* on the left in the image as guided by the question context, giving a correct prediction of *1*. However, the MAC model trained with biased general* attention slightly focuses on the target metallic object at the very beginning, and shifts its main attention to the *brown matte object* in the consecutive steps, which is not the final target the question is asking for, therefore it fails to make a correct prediction.

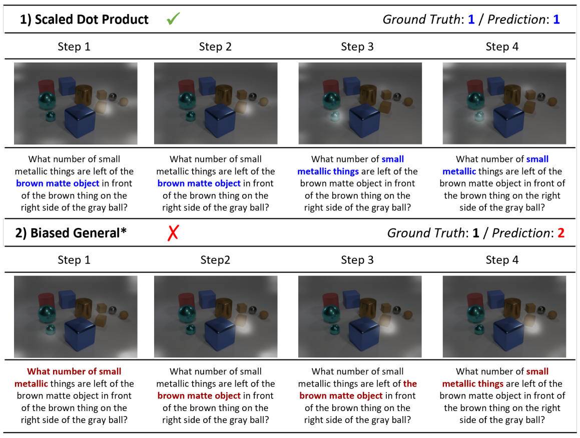

1) Scaled Dot Product ✓
Ground Truth: 1 / Prediction: 1
Step 1
Step 2
Step 3
Step 4
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
2) Biased General* ✗
Ground Truth: 1 / Prediction: 2
Step 1
Step2
Step 3
Step 4
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?
What number of small metallic things are left of the brown matte object in front of the brown thing on the right side of the gray ball?

Figure 5: Extra qualitative examples of VQA-CLEVR from the MAC trained by different attention alignment functions.

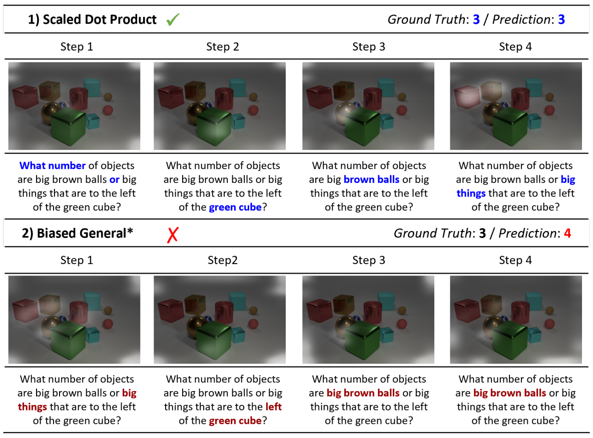

1) Scaled Dot Product ✓
Ground Truth: 3 / Prediction: 3
Step 1
Step 2
Step 3
Step 4
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?
2) Biased General* ✗
Ground Truth: 3 / Prediction: 4
Step 1
Step2
Step 3
Step 4
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?
What number of objects are big brown balls or big things that are to the left of the green cube?

Figure 6: Extra qualitative examples of VQA-CLEVR from the MAC trained by different attention alignment functions.

In Figure 6, a question *what number of objects are big brown balls or big things that are to the left of the green cube* is asked. MAC model using scaled dot product attention approaches this question by firstly attending to *what number* and *or* to understand the key concept that the question is asking for - the number of the union of two groups of objects. Then in the consecutive steps it focuses on the key objects *green cube*, *brown balls*, and remaining *big things* on the left of the green cube, so it can successfully give the correct answer *3*. However the model trained with biased general* attention focused on the *big things* before noting the condition *left of the green cube*, and failed to filter out irrelevant objects, giving a wrong prediction *4*.

In Figure 7, a question *are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects* is asked. MAC model using scaled dot product attention approaches this question by firstly attending to *same number* to count and compare relevant targets. Then it focuses on the key objects *purple metal object*, *brown rubber objects*, and *green blocks* on the

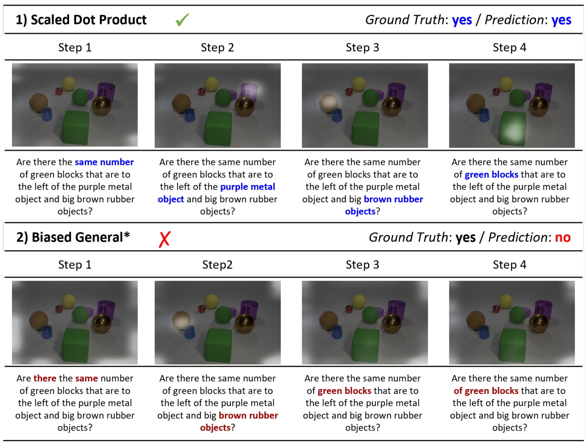

1) Scaled Dot Product ✓
Ground Truth: yes / Prediction: yes
Step 1
Step 2
Step 3
Step 4
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
2) Biased General* ✗
Ground Truth: yes / Prediction: no
Step 1
Step2
Step 3
Step 4
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?
Are there the same number of green blocks that are to the left of the purple metal object and big brown rubber objects?

Figure 7: Extra qualitative examples of VQA-CLEVR from the MAC trained by different attention alignment functions.

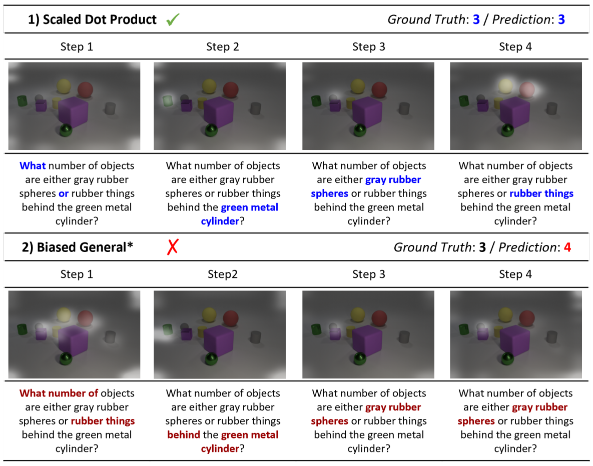

1) Scaled Dot Product ✓
Ground Truth: 3 / Prediction: 3
Step 1
Step 2
Step 3
Step 4
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
2) Biased General* ✗
Ground Truth: 3 / Prediction: 4
Step 1
Step2
Step 3
Step 4
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?
What number of objects are either gray rubber spheres or rubber things behind the green metal cylinder?

Figure 8: Extra qualitative examples of VQA-CLEVR from the MAC trained by different attention alignment functions.

left of the purple metal object in both question and the image, so it can successfully give the correct answer *yes*. However the model trained with biased general* attention focused on *green blocks* in the question in the last two steps but failed to find the target in the image, thus giving a wrong prediction *no*.

In Figure 8, a question *what number of objects are either gray rubber spheres or rubber things behind the green metal cylinder* is asked. MAC model using scaled dot product attention approaches this question by firstly attending to *what* and *or* in the question. Then it focuses on the relevant objects *green metal cylinder*, *gray rubber spheres*, and remaining *rubber things* in both question and the image, so it can successfully give the correct answer *3*. However the model trained with biased general* attention firstly focused on the *number of rubber things* before noting the condition *behind the green metal cylinder*, so it failed to filter out irrelevant objects, giving a wrong prediction *4*.

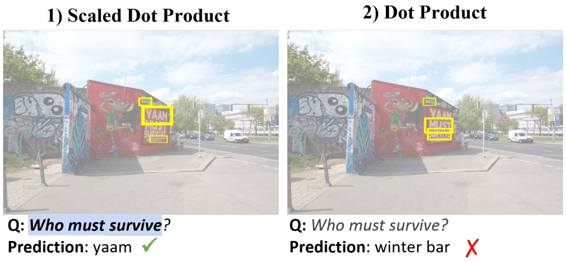


Figure 9: Extra qualitative examples of TVQA-TextVQA from the M4C trained by different attention alignment functions. Question words that receive attention weights greater than 0.01 are indicated in bold and coloured in blue.

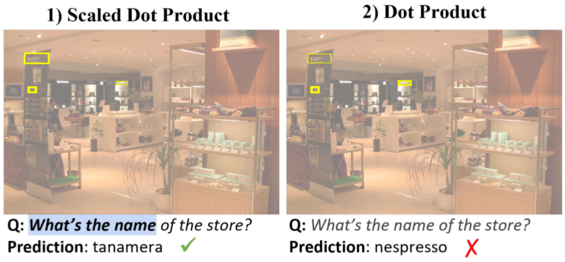


Figure 10: Extra qualitative examples of TVQA-TextVQA from the M4C trained by different attention alignment functions. Question words that receive attention weights greater than 0.001 are indicated in bold and coloured in blue.

A.4 Additional Qualitative Examples - TVQA

We also include more qualitative examples for M4C model in this section to show the difference between scaled dot product and dot product attention in the context of TVQA. In Figure 9, all the three words from the question *who must survive* received attention > 0.01 in the scaled dot product model, and the target OCR answer in the image received top attention among all OCR tokens. However the dot product model put much less attention on all question words, instead the OCR tokens for *must* and *survive* in the image were receiving top attention weights, followed by OCR token *winter* which is irrelevant to the question. Therefore scaled dot product model predicted correctly but dot product model did not.

In Figure 10, keywords *what's the name* from the question *what's the name of the store* received

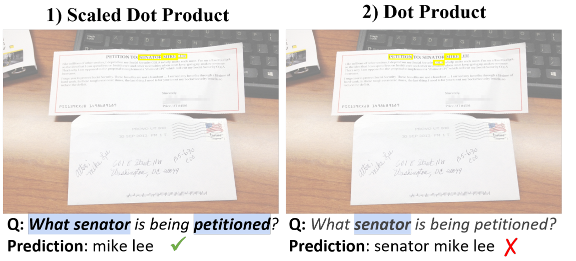

1) Scaled Dot Product
2) Dot Product
Q: What senator is being petitioned?
Prediction: mike lee ✓
Q: What senator is being petitioned?
Prediction: senator mike lee ✗

Figure 11: Extra qualitative examples of TVQA-TextVQA from the M4C trained by different attention alignment functions. Question words that receive attention weights greater than 0.001 are indicated in bold and coloured in blue.

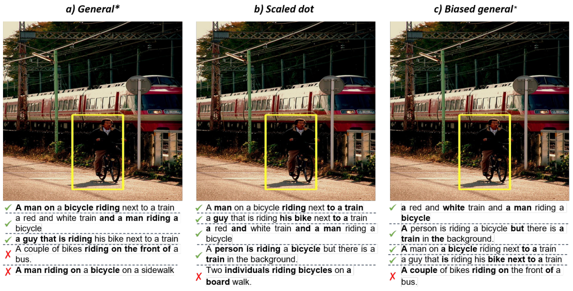

a) General*
b) Scaled dot
c) Biased general*
✓ A man on a bicycle riding next to a train
✓ a red and white train and a man riding a bicycle
✓ a guy that is riding his bike next to a train
✗ A couple of bikes riding on the front of a bus.
✗ A man riding on a bicycle on a sidewalk
✓ A man on a bicycle riding next to a train
✓ a guy that is riding his bike next to a train
✓ a red and white train and a man riding a bicycle
✓ A person is riding a bicycle but there is a train in the background.
✗ Two individuals riding bicycles on a board walk.
✓ a red and white train and a man riding a bicycle
✓ A person is riding a bicycle but there is a train in the background.
✓ A man on a bicycle riding next to a train
✓ a guy that is riding his bike next to a train
✗ A couple of bikes riding on the front of a bus.

Figure 12: Extra qualitative examples of T&I Match-MSCOCO from the SCAN trained by different alignment functions.

attention >0.001 in the scaled dot product model, similarly the dot product model put much less attention on all question words, and none of the question words received attention >0.001. Both models put most attention weights on the OCR token *gift* from the image, but scaled dot product managed to put more focus on the store name *tanamera* than the coffee brand name *nespresso*, which is the opposite case of the dot product model. Therefore scaled dot product model predicted correctly but dot product model did not.

In the example shown by Figure 11, there are lots of OCR tokens present in the image, making it more difficult to retrieve the correct answer tokens. As we can see from the picture, the model learned using dot product attention diverted its top attention to unrelated OCR token *cola*, and the top 3 OCR tokens receiving highest attention (*mike*, *cola* and *petition*) are not aligned with the predicted answer tokens (*senator mike lee*), while the model learned using scaled dot product attention put highest attention to expected or related OCR tokens that are aligned with the ground truth answers.

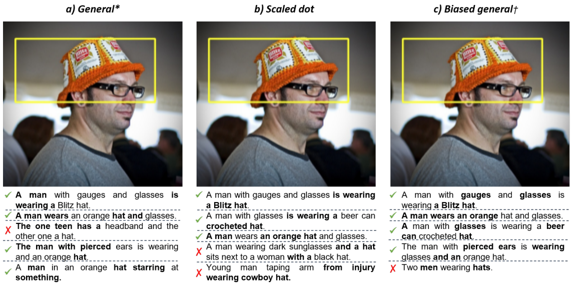

a) General*
b) Scaled dot
c) Biased general†
✓ A man with gauges and glasses is wearing a Blitz hat.
✓ A man wears an orange hat and glasses.
✗ The one teen has a headband and the other one a hat.
✓ The man with pierced ears is wearing and an orange hat.
✓ A man in an orange hat starring at something.
✓ A man with gauges and glasses is wearing a Blitz hat.
✓ A man with glasses is wearing a beer can crocheted hat.
✓ A man wears an orange hat and glasses.
✗ A man wearing dark sunglasses and a hat sits next to a woman with a black hat.
✗ Young man taping arm from injury wearing cowboy hat.
✓ A man with gauges and glasses is wearing a Blitz hat.
✓ A man wears an orange hat and glasses.
✓ A man with glasses is wearing a beer can crocheted hat.
✓ The man with pierced ears is wearing glasses and an orange hat.
✗ Two men wearing hats.

Figure 13: Extra qualitative examples of T&I Match-Flickr30k from the SCAN trained by different alignment functions.

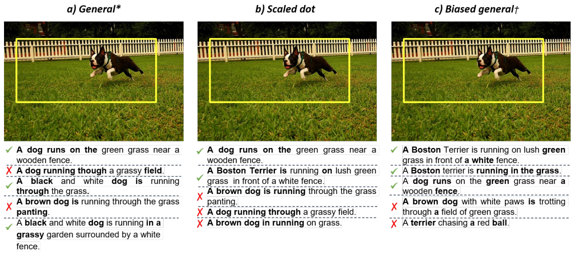

a) General*
b) Scaled dot
c) Biased general†
✓ A dog runs on the green grass near a wooden fence.
✗ A dog running though a grassy field.
✓ A black and white dog is running through the grass.
✗ A brown dog is running through the grass panting.
✓ A black and white dog is running in a grassy garden surrounded by a white fence.
✓ A dog runs on the green grass near a wooden fence.
✓ A Boston Terrier is running on lush green grass in front of a white fence.
✗ A brown dog is running through the grass panting.
✗ A dog running through a grassy field.
✗ A brown dog in running on grass.
✓ A Boston Terrier is running on lush green grass in front of a white fence.
✓ A Boston terrier is running in the grass.
✓ A dog runs on the green grass near a wooden fence.
✗ A brown dog with white paws is trotting through a field of green grass.
✗ A terrier chasing a red ball.

Figure 14: Extra qualitative examples of T&I Match-Flickr30k from the SCAN trained by different alignment functions.

A.5 Additional Qualitative Examples - T&I Match

In this section we visualize some examples for T&I Match models that show the attention received by retrieved captions with respect to the selected object region. In Figure 12 we can see that the best two models (i.e. models trained with general* attention or scaled dot product attention) can capture all key elements, *man, bicycle/bike, riding*, as the top attended words from the retrieved captions most of the time. However, the model trained using biased general* attention would capture at most one key element from each retrieved caption, and pay high attention to preposition, determinants or words related to other object regions. In the example shown by Figure 13, the model trained with scaled dot product attention can always capture the main object *hat* from all the retrieved captions, while the other two models sometimes fail to do so. In Figure 14, all three models sometimes wrongly recognise the dog's color (i.e. *brown dog* in wrongly retrieved captions). However, the best two models can retrieve the caption that is not in the ground truth list but also semantically matched to the given image (i.e. *A dog running through a grassy field*). The worst model trained with biased general† attention fails to do so, and it sometimes attends to objects from the caption that is not actually in the image (e.g. *red ball*).

A.6 Additional Qualitative Examples - T2I Gen

In this section we visualize and compare the best and the worst T2I Gen model. In Figure 15, the images generated by two models are highly similar but the worst model trained with activated general attention fails to attend to the key word *white*, so the bird it generated in the picture does not clearly have a white throat and chest.

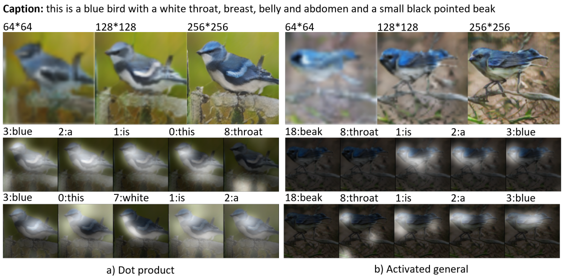


Figure 15: Extra qualitative examples of T2I Gen-CUB from the AttnGAN trained by different attention alignment functions.

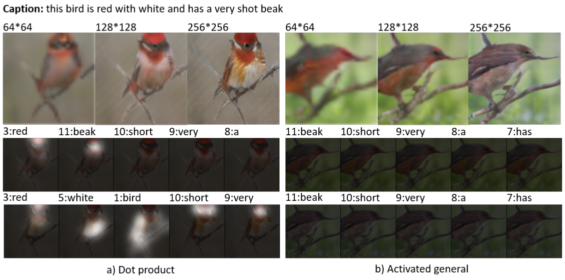


Figure 16: Extra qualitative examples of T2I Gen-CUB from the AttnGAN trained by different attention alignment functions.

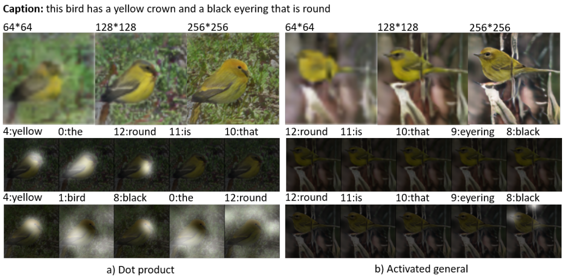


Figure 17: Extra qualitative examples of T2I Gen-CUB from the AttnGAN trained by different attention alignment functions.

In Figure 16 and Figure 17, the activation function used in the attention mechanism of the worst model makes it difficult to differentiate among the caption words when their attention weights are all very low. Therefore the model fails to attend to any useful facts in each attention layer, which makes it impossible to provide interpretability of model decision, despite generating an image that can roughly match the description in Figure 17. In Figure 16 the quality of the generated image is even worse - the feature of the bird does not match with the key phrases in the description (i.e. *red with white, short beak*).