Stacked Latent Attention for Multimodal Reasoning

Haoqi Fan
Facebook Research
1 Hacker Way
haoqifan@fb.com

Jiatong Zhou
Facebook Research
1 Hacker Way
jiatong@fb.com

Abstract

Attention has shown to be a pivotal development in deep learning and has been used for a multitude of multimodal learning tasks such as visual question answering and image captioning. In this work, we pinpoint the potential limitations to the design of a traditional attention model. We identify that 1) current attention mechanisms discard the latent information from intermediate reasoning, losing the positional information already captured by the attention heatmaps and 2) stacked attention, a common way to improve spatial reasoning, may have suboptimal performance because of the vanishing gradient problem. We introduce a novel attention architecture to address these problems, in which all spatial configuration information contained in the intermediate reasoning process is retained in a pathway of convolutional layers. We show that this new attention leads to substantial improvements in multiple multimodal reasoning tasks, including achieving single model performance without using external knowledge comparable to the state-of-the-art on the VQA dataset, as well as clear gains for the image captioning task.

1. Introduction

Attention has, in recent years, demonstrated to be a simple and effective mechanism for a neural network to "focus" on salient features of the input. Given an input state, attention allows the model to dynamically learn weights to indicate the importance of different parts of the input feature. It has been particularly successful for tasks requiring combined reasoning on multiple modalities such as visual question answering [15], image captioning [22], image-text matching [16], visual grounding [4] and others.

However, current attention mechanisms have several potential limitations which will impact performance. In order to learn which parts of the input to focus on, the attention mechanism reasons about the spatial information of the input, and "summarizes" the input state by computing a weighted sum to produce an embedding. Current attention models discard the latent spatial knowledge produced by the intermediate reasoning step, and only use the embedding output, representing the focus of the attention. This can potentially inhibit the model's spatial reasoning ability as by discarding the intermediate reasoning step, the model loses information pertaining to the position of the focal point. In order to improve the spatial reasoning ability of attention, there has also been research in the literature on multi-step reasoning. A common technique to do this is to stack attention layers, shown in the literature to improve performance by providing the model with a second chance to reason on the spatial dimension. This technique has seen varying degrees of success [15][16][10], and in this work, we show that one of the limitations for such models is that they are prone to the vanishing gradient problem, rendering the entire network ineffective.

To mitigate the above issues, in this paper we propose a stacked latent attention model, which can effectively capture the spatial knowledge computed in the attention mechanism and propagate this information through the network. Doing this also helps with stacking as the positional reasoning is provided to latter layers in the stack allowing for corrections to mistakes made in previous layers. Furthermore, the model is designed to mitigate the gradient vanishing problem with a pathway for the gradients to flow without being diluted by a softmax. This model is composed of attention units which possess the same input and output format as a traditional attention mechanism and can be used as a direct replacement in any task utilizing attention.

Building on top of this, we reinforce the technique of using attention for multimodal fusion, and propose a twin stream stacked latent attention model which is able to capture positional information from both textual and visual modalities. We provide a deep examination of this new model in the context of the VQA task, showing that we are able to achieve single model performance comparable to the state of the art without using any external knowledge. Additionally, we also explore the performance improvements of the stacked latent attention model over traditional attention models on another well developed multimodal learning

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00118

task: image captioning.

In this work, our contribution is three fold:

1) We pinpoint an overlooked limitation of attention models with respect to their ability for spatial reasoning and explore the issues of using stacked attention to alleviate this problem.

2) We propose a stacked latent attention model to tackle these problems which can be used as a direct replacement for current attention mechanisms. We further build on top of this to develop a twin stream stacked latent attention model for the fusion of knowledge in different modalities.

3) We demonstrated the effectiveness of the novel ideas in this paper on two tasks requiring multimodal learning: VQA and image captioning. For the VQA task we are able to achieve single model performance without using external knowledge comparable to the state of the art, and for image captioning, we show that a simple replacement of the attention mechanism to a SLA model can directly improve the baseline.

2. Related Work

In the multimodal understanding literature so far, attention has had a short but rich history. Beginning with [22], which developed attention to tackle the image captioning problem, attention has since gained popularity in a variety of different tasks, including visual question answering [20], image captioning [14], and others. With this increased adoption, a variety of attention models were developed to tackle problems of increasing difficulty including [25] which proposed a method combining both top-down and bottom-up approaches to extract richer information from semantic attributes detected from images and [24] which used attention on both the input image and encoder hidden states to capture the historical information of the LSTM memory cells. These works primarily focused on how to cast attention into the framework of a recurrent neural network to improve the capacity of the encoder-decoder framework, but did not investigate how to improve the attention itself. Another difficult problem for which attention mechanisms do well is the VQA task. Zhou et al. proposed a simple but elegant network in [28] which uses attention separately on the input image and text, and Teney et al. in [21], used features from the bottom up approach from [2], introduced many novel tricks to improve performance. All of these models derive the value from attention by utilizing the output embedding and discard the latent knowledge from the intermediate reasoning step.

To enhance the reasoning ability of attention for multimodal understanding, many works in the literature explored strategies employing multiple layers of attention. This was especially prevalent for the VQA task, where [15] proposed a Hierarchical Co-Attention model which can attend to image and question simultaneously with three attention layers. [23] was one of the first works to explore using a stacked attention architecture, although later work from Google in [10] reported only marginal improvements from stacking attention. More recent work done in [16] utilizes element-wise multiplication and residual connections to fuse information across modalities to be passed to the next attention unit. In these cases of multiple or stacked attention models, the intermediate spatial reasoning knowledge is again discarded. This can have a negative effect where latter layers of attention are particularly vulnerable to errors made in preceding layers. The Stacked Latent Attention model developed in this paper seeks to address these issues by propagating the latent positional information of the intermediate reasoning step through a series of convolutional layers, which acts as a natural pathway for the gradient to flow. In the VQA task, our model achieves state-of-the-art performance and demonstrates continued performance gains with three layers, in contrast to [16] where more than two stacked layers resulted in reduced performance. Moreover, the Stacked Latent Attention framework introduced in this paper can be modified to incorporate any one of these attention mechanisms.

Attention models employing more complex methods to compute correlation have also been proposed such as Multimodal Compact Bilinear Pooling [4], Low Rank Bilinear Pooling [11], MUTAN [3], and Multi-modal Factorized Bilinear Pooling[27]. These models provide a decent performance increase by using different types of bilinear pooling to provide richer representations of the image and text joint embedding than simpler linear models. However, these types of models are memory intensive and GPU inefficient due to the large dimensionality of the compacted/low rank outer product. In comparison, the Stacked Latent Attention model is more memory friendly as only convolutional layers are used and can still demonstrate strong performance.

3. Method

In this section, we first describe the formulation of the traditional attention mechanism and the stacked variant used to further spatial reasoning ability. Then we introduce our Stacked Latent Attention (SLA) model and detail its improvements. Finally, we build upon the SLA model to construct the Dual Stream Stacked Latent Attention model, designed to tackle the problem of multimodal joint learning on the Visual Question Answering [20] task.

3.1. Standard Attention Mechanism

While in the literature there exists many different methods and usages for attention, we can outline the general mechanism for attention as follows:

Given an input consisting of a set of K vectors of dimension D: $v = \{v_1, ..., v_K\}, v_i \in \mathbb{R}^D$ and an input state of size H: $h \in \mathbb{R}^H$, we can formulate the relative importance e_i of vector v_i to h as $e_i = f_{att}(v_i, h)$ where $f_{att}(v_i, h)$ is normally a two layer perceptron [4] [28] giving $f_{att}(v_i, h) = W_u\sigma(W_v v_i + W_h h)$ where $W_v \in \mathbb{R}^{D,m}$, $W_h \in \mathbb{R}^{H,m}$ and $W_u \in \mathbb{R}^{m,1}$, where m is the dimension of the perceptron's hidden layer. We can consider this two layer perceptron f_{att} as an intermediate reasoning step for the attention, and can be replaced by other models such as Multimodal Compact Bilinear Pooling [4], Low-Rank Bilinear Pooling [11], or other power approaches [27][3]. The attention weights α_i of each of the vectors v_i can be calculated by normalizing the relative importance e_i with a softmax as $\alpha_i = \frac{exp(e_i)}{\sum_k^K exp(e_k)}$. Finally, we generate the content vector of the attention as $z = \sum_i^K \alpha_i v_i$.

One of the limitations of this mechanism is that while the reasoning inside f_{att} contains a lot of useful latent spatial information: $s_i = \sigma(W_v v_i + W_h h)$, it is discarded after computing the content vector. This leads to the loss of rich location information and thus reduced ability for spatial reasoning. For example, given an image as input, the information pertaining to *where* the focus of the attention is, which is not saved in the content embedding z, is discarded.

3.2. Stacked Attention Model

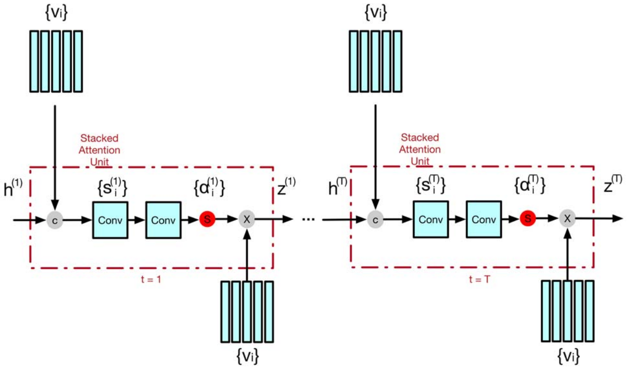

Figure 1. The main architecture of a traditional stacked attention model. Given the input $h^{(t)}$ and v_i, the attention model will generate normalized learnable weights $\alpha_i^{(t)}$, to conduct a weighted sum and generate the attention output $z^{(t)}$. We observe all activation functions are in the same pathway when stacking multiple attention units, meaning there will be T softmax layers in one pathway. We also see that inside the attention layer, the intermediate state has a spatial dimension and depth dimension, but when the output is generated, the dimension is reduced to d. This happens T times in a stack of T layers and becomes an information bottleneck. Here, the c in a gray circle is concatenate, s in a red circle is softmax, × in a grey circle is weighted sum.

In order to enhance the spatial reasoning ability of attention, there have been several attempts in the literature [23] [16] [15] of stacking different attention layers together. This works by using the content embedding $z^{(t)}$ of the last attention as the input $h^{(t)}$ to the next attention, with papers [16] [15] reporting gains for certain tasks. However there is some debate about this as [10] showed only marginal gains achieved when using the stacked attention strategy on the same tasks reported by [23].

The general architecture of a stacked attention model is described in Fig. 1. In the following formulations, we use the superscript notation to denote the layer of the attention. Given input vectors v_i, an initial hidden state $h^{(1)}$ and a stack of $t = T$ attention layers, we can describe attention as follows. The output of the $t'th$ attention layer is calculated as $z^{(t)} = \sum_i^K \alpha_i^{(t)} v_i^{(t)}$, where $\alpha_i^{(t)} = \frac{exp(f_{att}^{(t)}(v_i, h^{(t)}))}{\sum_k^K exp(f_{att}^{(t)}(v_k, h^{(t)}))}$. Here $f_{att}^{(t)}$ is the perceptron for the $t'th$ attention model. The spatial reasoning information on the $t'th$ attention layer is defined as $s_i^{(t)}$. For the next layer, we set $h^{(t+1)}$ to $z^{(t)}$. Finally, we can compute the final output of the stacked attention model as $z^{(T)}$.

When examining stacked attention, there are three important points that are often forgotten. First, only feeding the content embedding of the previous attention into the next is a potentially sub-optimal method for stacking attention. The latent positional information of the input contained in $s_i^{(t-1)}$ from the the previous attention is not considered at all in the next attention, which now has to perform the spatial reasoning from scratch. Additionally, carefully considering the pathway in Fig. 1, can lead to the observation that the dimensionality changes from $\mathbb{R}^{K,m}$ in $s_i^{(t)}$ to $\mathbb{R}^D$, then expands again in the next layer. It can be inferred that the $\mathbb{R}^D$ dimensionality becomes a bottleneck for the passage of information, limiting the performance of the network. Second, if the first attention is focused on the wrong position, the second attention will now only have the incorrect focus as input, which can heavily jeopardize performance. Finally, another observation in Fig. 1 shows that all activation functions and the softmax layer are on the same pathway. This can potentially cause the gradient to be diluted, making such a stacked architecture prone to gradient vanishing. We visualize the gradient in Fig. 5 to confirm the presence of gradient vanishing.

3.3. Stacked Latent Attention Model

To solve the issues identified above with both the traditional attention mechanism and the stacked variant, we propose the Stacked Latent Attention (SLA) model as shown in Fig. 2. Compared to the traditional stacked attention model introduced in Sec. 3.2, we introduced a conceptually simple change to allow each Stacked Latent Attention unit to utilize the spatial reasoning information from the previous unit. This allows for the following benefits: 1) By propagating the latent spatial reasoning information, the next attention layer can revise any errors from the previous layer. Furthermore the additional positional knowledge should help to improve the performance of the attention. 2) By designing the SLA model as described in the figure, we observe that

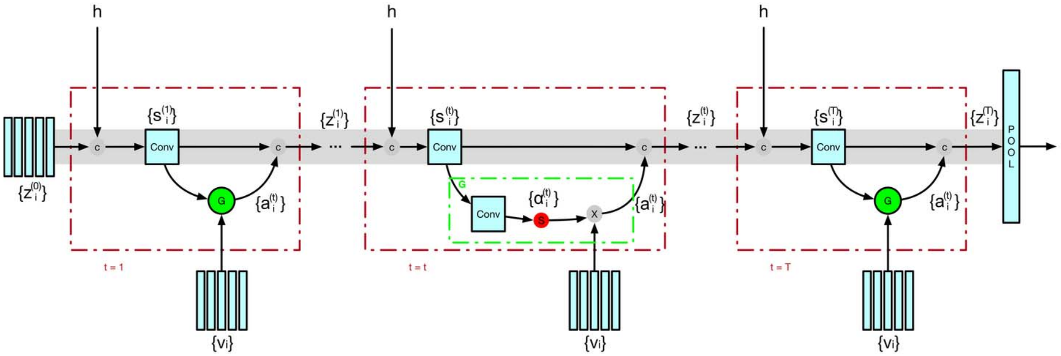

Figure 2. This figure illustrates the structure of the Stacked Latent Attention model. The model takes as input a hidden state h, initial spatial information $z_i^{(t)}$, input vector v_i. All knowledge flows through the main fully convolutional pathway, and the attention is learned in a separate path depicted in the dashed green box. Because the softmax is not in the main pathway, this model is able to greatly mitigate the gradient vanishing problem. The attention weights are generated as $\alpha_i^{(t)}$, and used in a weighted sum of the original input vectors v_i to generate the output attention embedding. This embedding conveys the global information of the entire input v_i, which is brought back to the main pathway. In the end, a spatial pooling is used to summarize the spatial information. For simplicity, we use G to represent the structure in the dotted green box for $T = 1, 3$.

the unit G containing the softmax is outside the main pathway. This helps to mitigate the gradient vanishing problem.

The SLA model can be defined as follows: given three inputs, v_i as the input vector, $h \in \mathbb{R}^H$ as the hidden state, and $z^{(t-1)} \in \mathbb{R}^{K,n^{(t-1)}}$ as the last SLA unit output, where $n^{(t)}$ is the embedding dimension on the $t'th$ Stacked Latent Attention unit, then, $f_{SLA}(v_i, h, z_i^{(t-1)}) = W_u^{(t)} s_i^{(t)}$ where $s_i^{(t)} = \sigma(W_v^{(t)} v_i + W_h^{(t)} h + W_z^{(t)} z^{(t-1)})$, $W_v^{(t)} \in \mathbb{R}^{D,m}$, $W_h^{(t)} \in \mathbb{R}^{H,m}$, $W_z^{(t)} \in \mathbb{R}^{n^{(t)},m}$ and $W_u \in \mathbb{R}^{m,1}$, here m is the dimension of the hidden layer and $n^{(t)}$ is the output depth dimension of the $t'th$ SLA unit. We can compute the normalized weights as $\alpha_i^{(t)} = \frac{exp(f_{SLA}^{(t)}(v_i,h,z_i^{(t-1)})}{\sum_k^K exp(f_{SLA}^{(t)}(v_i,h,z_i^{(t-1)})}$. Finally the output of the $t'th$ attention $z^{(t)}$ can be given by its K rows where the $i'th$ row is defined as $z_i^{(t)}$, the concatenation of $s_i^{(t)}$, the $i'th$ row of $s^{(T)}$, and $a^{(t)} = \sum_k^K \alpha_k^{(t)} v_k$.

The main pathway in SLA is fully convolutional, so at each point in the main pathway, there is always knowledge pertaining to each of the K positions of the input. This differs to the traditional attention model or stacked attention model where the output content embedding can not represent the location information of the input. Additionally, by using the spatial reasoning output $s_i^{(t)}$ to help predict the task, we can directly get the supervisory feedback information on the spatial information from the final loss without additional human labeled supervision as [18]. We concatenate the knowledge of the individual positions $s_i^{(t)}$ with the content embedding of the attention $\sum_k^K \alpha_k^{(t)} v_k$ to bring the spatial reasoning from a "global" perspective back to the main pathway. At the end of the attention chain, a pooling layer is used to summarize information on the spatial domain and produce a content vector as the final output.

Since we have introduced $z^{(t-1)}$ as input to the SLA unit, we are now given the opportunity to introduce positional bias into the initial attention layer. We achieve this through bootstraping z^0 with the image input v concatenated with a positional bias b_{pos}. To generate b_{pos}, we can simply feed the first hidden state h into a fully differentiable model. This is intuitively understandable, for example in the VQA task with the question as "what is the color of the sky", the word embedding can be used to generate an initial spatial heatmap with high activations near the top of the image where one would typically find the sky. We visualize this initial heatmap in the experiment section 8 to very interesting results.

3.4. Twin Stream Stacked Latent Attention Network

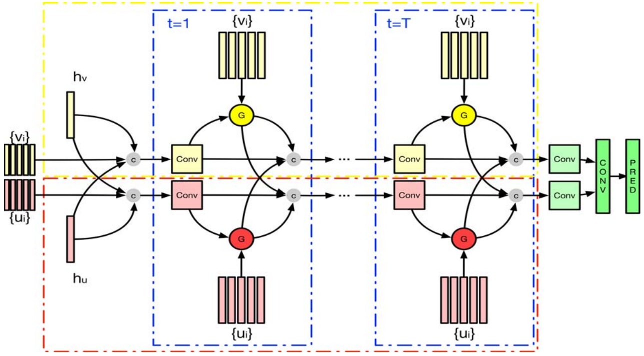

Figure 3. This figure shows the main framework of the Twin Stream SLA model. Each stream is a Stacked Latent Attention network. In order to fuse the information from multiple modalities, the weighted sum attention output of one stream is concatenated with the attention outputs of the other stream. h_v and h_u are the initial hidden states in the visual and textual streams generated by pooling and a LSTM respectively. The final output from both streams are fused together by a two layer neural network. For simplicity, we omit the details in G which can be seen in Fig. 2.

By building upon the SLA model, we construct the twin stream Stacked Latent Attention model which is designed to apply the same ideas to merge knowledge from both the visual and textual modalities. As the name suggests, this model introduces two streams, denoted by the superscript vis to represent the visual stream and $text$ to represent the textual stream. Additionally we define the spatial visual input as v_i and the temporal textual input as u_i.

For the visual input, we extract the feature representation $\{v_i, ...v_K\}$ for an input image from a deep neural network, where K is the number of regions. For the textual input we have embeddings for the words as $\{u_i, ..., u_L\}$, where L is the number of words. Additionally, we use $h^{(0),vis}$ and $h^{(0),text}$ to represent the initial hidden state of the visual and textual streams respectively. $h^{(0),text}$ is initialized by using a bi-directional LSTM on top of $\{u_i, ..., u_N\}$ so that we get $h^{(0),text} = bi - LSTM(\{u_i, ..., u_L\})$, and $h^{(0),vis}$ can be obtained with a simple spatial pooling on $\{v_i, ...v_K\}$.

As show in Fig. 3, the key idea of the Twin Stream SLA is that we can allow for the interchange of information between the two modalities by simply concatenating the weighted sum attention output of one stream to the attention outputs of the other stream. More specifically we can use the attention embedding from the textual stream as a hidden state input for the next attention layer of the visual stream: $f^{vis}_{SLA}(v_i, h^{(t),text}, z_i^{(t-1),vis})$ and symmetrically, the attention embedding of the visual stream is used as a hidden state input for the next attention in the textual stream: $f^{text}_{SLA}(u_i, h^{(t),vis}, z_i^{(t-1),text})$. By doing this we reinforce the reasoning of one stream with the knowledge of the other modality. After T stacks, the output from both streams are fed through a separate convolutional and pooling layer and then concatenated, passing through one FC layer then projected to a C dimensional space, where C is the number of classes.

Moreover, we augment the spatial stream by adding a positional bias to the input. We use h^{text} as the seed for the positional bias by feeding it into one fully connected layer, to obtain b_{pos}, which is then concatenated with v_i to form the input of the spatial stream. We can regard this as the positional bias generated by the word embedding and show in Fig. 8 that it is especially useful for the VQA task. An intuitive example to demonstrate this is, given the question of "what is the color of the ground", we can bias the model about where to look even before it sees the image.

4. Experiments

In this section, we detail the deep investigation of the SLA mechanism for the VQA task in Section 4.1 and also evaluate its benefits for the Image Captioning task in Section 4.2.

4.1. Twin Stream Stacked Latent Attention Model For VQA

We choose the VQA task on which to perform several experiments to evaluate the effectiveness of the twin stream Stacked Latent Attention model. VQA is one of the best tasks to explore the capability of new types of attention as a significant number of VQA research [21][27][4][3][15][16][23][18] applies some type of attention. The role of attention in the VQA problem can be expressed as: given an input image and a sequence of words forming a question, find the optimal set of image regions and words to answer the question.

4.1.1 Dataset and Evaluation Metric

We analyze the performance of the twin stream Stacked Latent Attention model on the VQA 2.0 Dataset [6] consisting of 204K images and 614K free-form natural language questions. Each image is associated with three questions, and each question is labeled with ten answers by human annotators. The dataset is typically divided into four splits: train (80K images), val (40K images), test (81K images). For this task, we follow the evaluation metric used in [20] as

$$Acc(\text{ans}) = min\{\frac{\#human\ that\ labeled\ ans}{3}, 1\}$$

where ans is a predicted answer.

4.1.2 Experimental Setup

For VQA dataset, the experiments are set up as follows. The input images are scaled while preserving aspect ratio, as suggested in [28], and center cropped and scale to a dimension of 448×448. Then, the image features are extracted using a pretrained 200 layer ResNet [8] model, specifically we use the last layer before the average pooling layer to obtain a $14 \times 14 \times 2048$ tensor and then perform L2 normalization for each $1 \times 1 \times 2048$ spatial feature. On the question side, the words are tokenized and projected to a 300 dimensional space. This embedding is then passed to a bi-directional LSTM with a hidden size of 512. We set a maximum length for the question, selecting the first 26 words.

The model is optimized with the ADAM optimizer [12] with a batch size of 800 and trained on 8 GPUs. The ADAM hyperparameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and all model parameteres are initialized as suggested by Glorot et al. in [5]. The Visual Genome dataset [19] is not used for training. All the numbers reported are of a single model without ensembling.

4.1.3 Comparison to State-of-the-Art

Recently there have been many different approaches which use different types of external knowledge to augment the

Single Model Performance	Test-dev				Test-standard			
	All	Y/N	Num	Other	All	Y/N	Num	Other
Prior (most common answer in training set) [6]	-	-	-	-	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [6]	-	-	-	-	44.26	67.01	31.55	27.37
Deeper LSTM Q norm [6]	-	-	-	-	54.22	73.46	35.18	41.83
Soft Loss Function [7]	60.4	71.9	39.0	54.6	-	-	-	-
Human-like Attetnion [18]	61.99	78.54	37.94	53.38	-	-	-	-
VQA Challenge Winner 2017 [21]	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Multimodal Compact Bilinear Pooling [4] VQA Challenge Winner 2016	-	-	-	-	62.27	78.82	38.28	53.36
MFB [27]	64.98	-	-	-	-	-	-	-
MFH [27]	65.80	-	-	-	-	-	-	-
VQA Challenge Winner 2017 [21] with bottom-up attention *	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
Dual Recurrent Attention Units [1] with FRCNN Feature *	66.45	82.85	44.78	57.4	66.85	83.35	44.37	57.63
Our Single Model	63.89	79.95	40.35	55.86	64.06	80.01	40.63	55.82

Table 1. Comparison of our best single model with competing single model methods. * means the model is trained with region-specific features which can be regarded as extra knowledge.

visual question answering model. For example [18] manually annotated a human-like attention heatmap as a supervisory signal to train the attention model. [21] and [1] used external knowledge like Fast(er) R-CNN, showing significant improvements on the VQA task. [27] [4] [3] used visual Genome [19] to augment the VQA dataset. [27] [3] used skipthought vectors [13] to augment the textual feature while [4] [21] used Glove [17]. Others [4] [21] [27] [3] ensemble from 3 to 20 different models to boost performance. The many different configurations make it hard to have a standardized setup on which to fairly compare results. For this work, we report our number without any external knowledge either by augmenting the training set. A comparison to state-of-the-art VQA systems is presented in Table 1. We split the table into 3 sections, in the first section, we report all the existing numbers we can find for non-ensembled models that aren't using additional knowledge outside of the VQA dataset. In the second section, we report some noteworthy numbers in VQA which do use external knowledge and in the final section we report our result. It can be seen in these results that the twin stream Stacked Latent Attention model can achieve performance comparable to the state-of-the-art on the test-dev and test-standard dataset.

Furthermore, we also report our number of 65.3 on VQA 1.0, surpassing the performance of previous state-of-the-art results: 64.6 for MLAN[26], 64.3 for Dual Attention[16], 64.2 for MCBP[4], 61.8 for HieCoAtt[15], and 61.02 for MUTAN [3].

4.1.4 Analysis

Going beyond the quantitative results, we are interested in exploring the properties of the Stacked Latent Attention model, and why it can outperform traditional attention models or stacked attention models. In this section, we analyze the different properties of the model and reason about effects and contributions of each part.

First we inspect the performance delta after varying different hyperparameters of the twin stream Stacked Latent Attention model. For the main convolutional pathway, we compared hidden state sizes of 256, 512, and 1024. For a fair comparison, we did not try a size greater than 1024 as the model would not fit into GPU memory without lowering the batch size. We also analyze the effectiveness of different numbers of stacked layers T for $T = 1, 2, 3$. Here, we note an interesting finding that although [16] reports stacking 2 attention layers results in the best performance, [10] reports poor performance when stacking 3 layers and [4] reports that no performance gains emerge even by stacking 2 layers, in our model, as demonstrated in Fig. 4, we can clearly see that stacking more layers results in better performance. This is in line with our hypothesis that by exposing the latent positional information throughout the whole pathway and mitigating the gradient vanishing problem, stacking attention is generally beneficial. We observe that the best performance can be achieved by stacking 3 attention layers, although we did not try with more layers since we could not keep the same batch size.

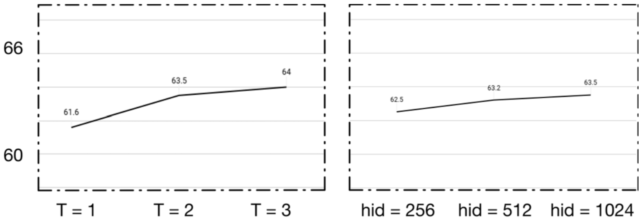

Figure 4. We compare the accuracy of the twin stream Stacked Latent Attention model with different hyperparameter settings. The figure on the left shows performance for $T = 1, 2, 3$, and the figure on the right shows performance for hidden state sizes of $256, 512, 1024$ for $T = 2$. We find improving performance with more layers, even for a stack of 3.

To dive deeper into why previous work was not able to successfully stack more than two layers of attention while in our work, we see that more layers give better performance, we visualize the distribution of gradient magnitude at each layer of the attention stack for the visual stream. This is shown in of Fig. 5 for both a baseline traditional stacked attention model as well as for the SLA model, with a stack

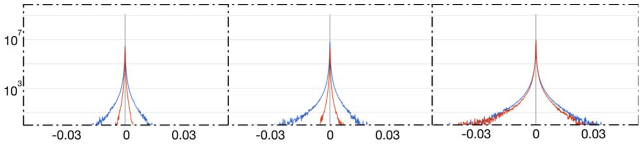

Figure 5. This figure illustrates the change in magnitude of the gradients at different layers of attention in the network for the visual stream on the VQA task. The leftmost figure depicts the gradient of the earliest attention and the rightmost figure depicts the gradient of the latest attention. The red curve is the baseline traditional stacked attention model while the blue curve is our SLA model. We observe that for layers closer to the output, the distribution has a large variance for both the standard stacked model and our SLA model, however as we move from the output layer towards the input, the gradient magnitude distribution peaks tightly around zero with low variance for the traditional stacked attention model. This suggests that our model is able to mitigate the vanishing gradient problem.

size $T = 3$. We fixed all other configurations of the experiment to ensure a fair comparison. In the figure, the x-axis denotes the gradient magnitude and the y-axis is the log frequency, the left, middle and right figures represent the histogram of the first, second and third attention layers respectively and the red curve represents the baseline model while the blue curve represents the SLA model. We observe that all of the gradient distributions are centered around 0. The interesting feature of this figure is that in the last attention, the one closest to the loss, both models have a similar distribution, however, as we move closer to the input the distribution of the gradient magnitude of the traditional stacked model peaks at 0 with very small variance. In contrast to the SLA model, we see continued high variance even at the initial layers indicating that this model does not suffer from gradient vanishing.

The reason for this can be seen in the difference of architecture between the traditional stacked model shown in Fig. 3.2 and the SLA model show in Fig. 3.3. In the traditional model, we observe that the gradient vanishes because all of the activations, including the softmax layer, are in the same pathway, so when the gradient backpropagates, it gets diluted by the softmax. In contrast, the SLA model is designed such that the main pathway is softmax free, with the softmax activations out of the main pathway. This allows the model to avoid falling prey to the vanishing gradient problem and greatly enhances the smooth flow of the gradient to the beginning of the network, resulting in the ability to stack more attention layers.

We perform further analysis to support this point by plotting the activations of the attention heatmap between the baseline model and the SLA model. The hypothesis is that, the more performant attention model should have sparser and stronger activations leading to a sharper attention. Fig. 6 plots the distribution of the strength of the activations. The x-axis denotes the magnitude of the activation and the y-axis plots the log frequency. The blue and red curves represent the SLA model and traditional stacked attention respectively. The leftmost figure depicts the first attention layer, while the middle figure depicts the middle attention layer and the rightmost figure depicts the last attention layer. It is clear from this figure that there are significantly more activations of a greater magnitude for the SLA model than the traditional stacked attention leading to a sharper attention that is able to more smoothly propagate the discriminatory supervision signal from the loss. In both cases, each subsequent layer has stronger activations than the previous layer which is within expectations since it is closer to the loss.

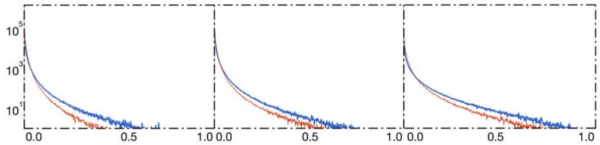

Figure 6. This is the visualization for the distribution of attention weights α_i for different timesteps $T = 1, 2, 3$ from left to right. The x-axis is the magnitude and the y-axis is the logged frequency. The red curve represents the standard stacked attention model, and the blue curve represents the SLA model. We observe that the SLA model generates a sharper attention than the normal stacked attention model, especially for the initial layer because the design of the SLA model can facilitate the smooth backpropagation of the gradient, giving the model a strong supervisory signal.

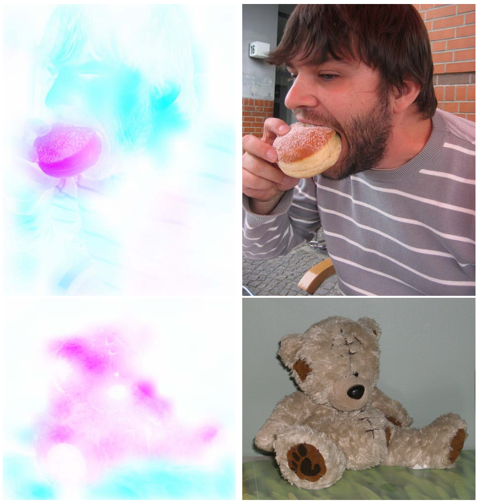

Figure 7. The blue shade is the visualization of the attention from the first step of inference and red shade shows the second step. From the visualization, the reasoning process becomes clear. With a question of ”what is the man eating?”, the spatially augmented attention can check for the man first, and then look at the food in the man’s mouth.

Fig. 7 presents some qualitative insight into the activations of the attention with a visualization of the image attention. The reasoning steps are clearly illustrated, with the blue shade representing the first attention unit and the red shade representing the last. In the first example, the ques-

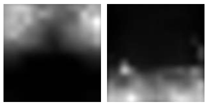

Figure 8. This figure illustrates the location bias generated by the word embedding. The left image represents the positional bias given “sky”, while the right image represents the spatial bias given “ground”.

tion is ”what is the man eating?”, and the first attention layer is focused on the man, corresponding to the word “man” in the question. The last attention layer can be seen to be focused on the object in the man’s hand producing the answer to the question: “donut”. Similarly, for the second example, the question is ”what is on the ground?”, and the first attention layer highlights the floor while the last attention is on the teddy bear.

	Y/N	Num	Other	All
Baseline T=2, hid=256	78.61	43.00	53.57	61.59
Baseline + location Bias	**80.15**	**43.97**	**54.54**	**62.77**

Table 2. Comparison of performance when adding initial positional bias to the baseline Twin Stream Stacked Latent Attention model with $T = 1$ and hidden state dimension of 256. A clear improvement can be observed.

Additionally, we also investigate the properties of the positional bias b_{pos} introduced in Section 3.3, we introduced the idea of using the word embedding to bootstrap the initial positional bias b_{pos}. The method of doing this is to pass the initial question to a bi-directional LSTM to generate an embedding which goes through a non-linear projection to generate b_{pos}. By visualizing the properties of this initial state in Fig. 8 we show that it contains some semantic meaning that is beneficial to the model. This figure is created by using a trained SLA model and directly visualizing b_{pos} given the separate questions of ”what is in the sky”, and ”what is the color of the ground”. The left figure shows that the upper part of b_{pos} activates given the question ”what is in the sky” and the right figure shows activation of the bottom half of the image given the question ”what is the color of the ground”. This demonstrates that the model can utilize the positional information learned from the word embedding. Table 2 compares the performance of the model with and without initializing b_{pos} with this positional bias. We see that the baseline model without this bias achieves an accuracy of 61.59 while after adding the positional bias we see an accuracy of 62.77, a gain of 1.18.

4.2. Stacked Latent Attention Model for Image Caption

In order to prove the generalizability of our model, in this section we explore its performance on the image captioning task. It is another task which investigates the fusion of vision and text for multimodal reasoning where attention models have demonstrated large impact [24][25][22][9]. For these experiments, we use the same framework as used in [24] and regard it as the baseline. The only change we make to this model is to replace the traditional attention mechanism with our SLA model with $T = 3$. We also compare with the classic encoder/decoder models described in [22].

The models are evaluated on the MSCOCO dataset [14] containing 123000 images, each with at least 5 captions. The dataset is split in the same way as in [24][25][22][9], with 5000 images reserved for the dev and test set each and the remainder used for training. As part of the preprocessing, we retain only alphabetic characters in the captions and convert all letters to lowercase. The captions are tokenized using white space. We only consider words occurring 5 or more times and replace the remainder with the UNK token giving a vocabulary of 9520 words. Captions are truncated to a max length of 30 tokens as we use the same hyperparameters as [24].

Table 3 reports the BLEU-4, METEOR and CIDEr scores for these experiments. They are produced using the official MSCOCO evaluation scripts. The results show that by simply switching to a SLA model for attention, we obtain a substantial gain on this task.

	BLEU-4	METEOR	CIDEr
Encoder-Decoder [22]	0.278	0.229	0.840
ReviewNet[24]	0.290	0.237	0.886
ReviewNet + SLA	0.300	0.253	0.908

Table 3. Performance of different model variants on the MSCOCO dataset. Results are obtained with a single model using VGGNet.

The image captioning task requires a detailed understanding about every section of the image. We can see that the SLA model is able to bring a performance increase of 0.01 on BLEU-4, 0.016 on METEOR and 0.022 on CIDEr compared to the ReviewNet. This can be attributed to two reasons, first, the multiple reasoning layers in the SLA unit provides a more detailed and accurate attention and second, the additional positional information provided by the SLA model can be used to generate a more precise description.

5. Conclusion

In this paper, we identify and explore the limitations of traditional attention models when conducting spatial reasoning, and the tendency for gradient vanishing when employing a stacked architecture. To tackle these problems, we propose the Stacked Latent Attention model which can be used as a direct replacement for current attention mechanisms and build upon this to develop the Twin Stream Stacked Latent Attention model for the Visual Question Answering task. With these models, we are able to achieve state-of-the-art performance for VQA and demonstrate clear improvements for the image captioning task.

References

[1] O. Ahmed and W. Samek. Dual recurrent attention units for visual question answering. *arXiv*, 1802.00209, 2018. 6

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention. *arXiv*, 1707.07998, 2017. 2

[3] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. *arXiv*, 1705.06676, 2017. 2, 3, 5, 6

[4] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 1, 2, 3, 5, 6

[5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249-256, 2010. 5

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arxiv*, 1612.00837, 2016. 5, 6

[7] I. Ilievski and J. Feng. A simple loss function for improving the convergence and accuracy of visual question answering models. *arXiv*, 1708.00584, 2017. 6

[8] H. Kaiming, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 5

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128-3137, 2015. 8

[10] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv*, 1704.03162, 2017. 1, 2, 3, 6

[11] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv*, 1610.04325, 2016. 2, 3

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 1412.6980, 2014. 5

[13] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, , and S. Fidler. Skip-thought vectors. *In Advances in neural information processing systems*, 32943302, 2015. 6

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, , and C. L. Zitnick. Microsoft coco: Common objects in context. *In Computer VisionECCV*, pages pp. 740–755, 2014. 2, 8

[15] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016. 1, 2, 3, 5, 6

[16] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *CoRR*, 2016. 1, 2, 3, 5, 6

[17] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing*, 1532-1543, 2014. 6

[18] T. Qian, J. Dong, and D. Xu. Exploring human-like attention supervision in visual question answering. *arXiv*, 1709.06308, 2017. 4, 5, 6

[19] Ranjay, Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, and S. C. et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 32-73, 2017. 5, 6

[20] A. Stanislaw, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. *In Proceedings of the IEEE International Conference on Computer Vision*, pages pp. 2425–2433, 2015. 2, 5

[21] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv*, 1708.02711, 2017. 2, 5, 6

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, 2048-2057, 2015. 1, 2, 8

[23] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. *In Computer Vision and Pattern Recognition*, 2016. 2, 3, 5

[24] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. *In Advances in Neural Information Processing Systems*, 2361-2369, 2016. 2, 8

[25] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4651-4659, 2016. 2, 8

[26] D. Yu, J. Fu, T. Mei, and Y. Rui. Multi-level attention networks for visual question answering. *In Conf. on Computer Vision and Pattern Recognition*, 2017. 6

[27] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *arXiv*, 1708.01471, 2017. 2, 3, 5, 6

[28] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015. 2, 3, 5