arXiv:1606.01847v2 [cs.CV] 23 Jun 2016

Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding

Akira Fukui*[1,2] **Dong Huk Park***[1] **Daylen Yang***[1]
Anna Rohrbach*[1,3] **Trevor Darrell**[1] **Marcus Rohrbach**[1]
[1]UC Berkeley EECS, CA, United States
[2]Sony Corp., Tokyo, Japan
[3]Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

Modeling textual or visual information with vector representations trained from large language or visual datasets has been successfully explored in recent years. However, tasks such as visual question answering require combining these vector representations with each other. Approaches to multimodal pooling include element-wise multiplication or addition, as well as concatenation of the visual and textual representations. We hypothesize that these methods are not as expressive as an outer product of the visual and textual vectors. As the outer product is typically infeasible due to its high dimensionality, we instead propose utilizing Multimodal Compact Bilinear pooling (MCB) to efficiently and expressively combine multimodal features. We extensively evaluate MCB on the visual question answering and grounding tasks. We consistently show the benefit of MCB over ablations without MCB. For visual question answering, we present an architecture which uses MCB twice, once for predicting attention over spatial features and again to combine the attended representation with the question representation. This model outperforms the state-of-the-art on the Visual7W dataset and the VQA challenge.

1 Introduction

Representation learning for text and images has been extensively studied in recent years. Recurrent neural networks (RNNs) are often used to represent sentences or phrases (Sutskever et al., 2014; Kiros et al., 2015), and convolutional neural networks (CNNs) are often used to represent images (Donahue et al., 2013; He et al., 2015). For tasks such as visual question answering (VQA), most approaches require joining the representation of both modalities. For combining the two vector representations (multimodal pooling), current approaches in VQA or grounding rely on concatenating vectors or applying the element-wise addition or multiplication. While this generates a joint representation, it might not be expressive enough to fully capture the complex associations between the two different modalities.

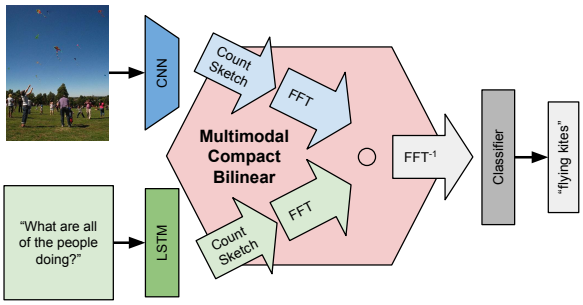

Figure 1: Multimodal Compact Bilinear Models for visual question answering.

In this paper, we propose to rely on Multimodal Compact Bilinear pooling (MCB) to get a joint representation. Bilinear pooling computes the outer product between two vectors, which allows, in contrast to element-wise multiplication, a multiplicative interaction between all elements of both vectors. Bilinear pooling models (Tenenbaum and Freeman, 2000) have recently been shown to be beneficial for fine-grained classification for vision only tasks (Lin et al., 2015). However, given their high dimensional-

* indicates equal contribution

ity (n^2), bilinear pooling has so far not been widely used. In this paper, we adopt the idea from Gao et al. (2016) which shows how to efficiently compress bilinear pooling for a single modality. In this work, we discuss and extensively evaluate the extension to the multimodal case for text and visual modalities. As shown in Figure 1, Multimodal Compact Bilinear pooling (MCB) is approximated by randomly projecting the image and text representations to a higher dimensional space (using Count Sketch (Charikar et al., 2002)) and then convolving both vectors efficiently by using element-wise multiplication in Fast Fourier Transform (FFT) space. We use MCB to predict answers for the VQA task and locations for the visual grounding task. For open-ended question answering, we present an architecture for VQA which uses MCB twice, once to predict spatial attention and the second time to predict the answer. For multiple-choice question answering we introduce a third MCB to relate the encoded answer to the question-image space. Additionally, we discuss the benefit of multiple attention maps and additional training data for the VQA task. To summarize, MCB is evaluated on two tasks, four datasets, and with a diverse set of ablations and comparisons to the state-of-the-art.

2 Related Work

Multimodal pooling. Current approaches to multimodal pooling involve element-wise operations or vector concatenation. In the visual question answering domain, a number of models have been proposed. Simpler models such as iBOWIMG baseline (Zhou et al., 2015) use concatenation and fully connected layers to combine the image and question modalities. Stacked Attention Networks (Yang et al., 2015) and Spatial Memory Networks (Xu et al., 2015) use LSTMs or extract soft-attention on the image features, but ultimately use element-wise product or element-wise sums to merge modalities. D-NMN (Andreas et al., 2016a) introduced REINFORCE to dynamically create a network and use element-wise multiplication to joint attentions, and elementwise sum to predict answers. Dynamic Memory Networks (DMN) (Xiong et al., 2016) pool image and question with element-wise multiplication and addition, attending to part of image and question with a Episodic Memory Module (Kumar et al., 2015). DPPnet (Noh et al., 2015) create a Parameter Prediction Network which learns to predict the parameters of the second last visual recognition layer dynamically from the question. Lu et al. (2016) recently proposed a model that extracts multiple co-attentions on the image and question and combines the co-attentions in a hierarchical manner using all of element-wise sums, concatenation, and fully connected layers.

For the visual grounding task Rohrbach et al. (2016) propose an approach where language phrase embedding are concatenated with the visual features in order to predict the attention weights over multiple bounding box proposals. Similarly, Hu et al. (2016a) concatenate phrase embeddings with visual features at different spatial locations to obtain a segmentation.

Bilinear pooling. Bilinear pooling has been applied to the fine-grained visual recognition task. Lin et al. (2015) use two CNNs to extract features from an image and combine the resulting vectors using an outer product, which is fully connected to an output layer. Gao et al. (2016) address the space and time complexity of bilinear features by viewing the bilinear transformation as a polynomial kernel. Pham and Pagh (2013) describes a method to approximate the polynomial kernel using Count Sketches and convolutions.

Joint multimodal embeddings. In order to model similarity between two modalities many prior works have learned joint multimodal spaces, or embeddings. Some of such embeddings are based on Canonical Correlation Analysis (Hardoon et al., 2004) e.g. (Gong et al., 2014; Klein et al., 2015; Plummer et al., 2015), linear models with ranking loss (Frome et al., 2013; Karpathy and Fei-Fei, 2015; Socher et al., 2014; Weston et al., 2011) or non-linear deep learning models (Kiros et al., 2014; Mao et al., 2015; Ngiam et al., 2011). Our multimodal compact bilinear pooling can be seen as a complementary operation that allows to capture different interactions between two modalities more expressively than e.g. concatenation. Consequently many embedding learning approaches could benefit from incorporating such interactions.

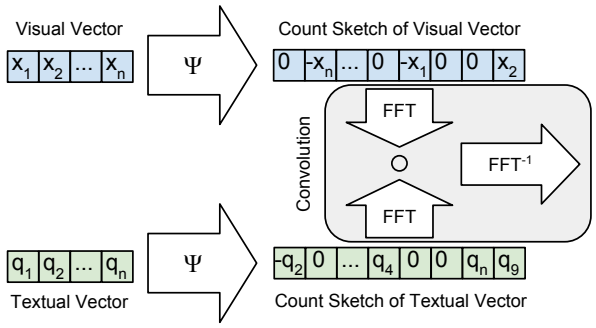

Figure 2: Multimodal Compact Bilinear Pooling (MCB)

3 Multimodal Compact Bilinear Pooling for Visual and Textual Embeddings

For the task of visual question answering (VQA) or visual grounding, we have to predict the most likely answer or location $\hat{a}$ for a given image $\mathbf{x}$ and question or phrase $\mathbf{q}$. This can be formulated as

$$\hat{a} = \underset{a \in A}{\operatorname{argmax}}\, p(a|\mathbf{x}, \mathbf{q}; \theta) \tag{1}$$

with parameters θ and the set of answers or locations A. For an image embedding $x = \Xi(\mathbf{x})$ (i.e. a CNN) and question embedding $q = \Omega(\mathbf{q})$ (i.e. an LSTM), we are interested in getting a good joint representation by pooling both representations. With a multimodal pooling $\Phi(x, q)$ that encodes the relationship between x and q well, it becomes easier to learn a classifier for Equation (1).

In this section, we first discuss our multimodal pooling Φ for combining representations from different modalities into a single representation (Sec. 3.1) and then detail our architectures for VQA (Sec. 3.2) and visual grounding (Sec. 3.3), further explaining how we predict $\hat{a}$ with the given image representation Ξ and text representation Ω.

3.1 Multimodal Compact Bilinear Pooling (MCB)

Bilinear models (Tenenbaum and Freeman, 2000) take the outer product of two vectors $x \in \mathbb{R}^{n_1}$ and $q \in \mathbb{R}^{n_2}$ and learn a model W (here linear), i.e. $z = W\,[x \otimes q]$, where $\otimes$ denotes the outer product (xq^T) and $[\,]$ denotes linearizing the matrix in a vector. As discussed in the introduction, bilinear pooling is interesting because it allows all elements of both vectors to interact with each other in a multiplicative way. However, high dimensional representation

Algorithm 1 Multimodal Compact Bilinear

input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
procedure MCB(v_1, v_2)
  for $i \leftarrow 1 \ldots 2$ **do**
    if h_i, s_i uninitialized **then**
      for $k \leftarrow 1 \ldots n_i$ **do**
        sample $h_i[k]$ from $\{1, \ldots, c\}$
        sample $s_i[k]$ from $\{1, -1\}$
    $v'_i \leftarrow \Psi(v_i, h_i, s_i)$
  $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
  return Φ
procedure $\Psi(v, h, s)$
  $y \leftarrow [0, \ldots, 0]$
  for $j \leftarrow 1 \ldots n$ **do**
    $y[h[j]] \leftarrow y[h[j]] + s[j] \cdot v[j]$
  return y

(i.e. when n_1 and n_2 are large) lead to an infeasible number of parameters to learn in W. For example, we use $n_1 = n_2 = 2048$ and $z \in \mathbb{R}^{3000}$ for VQA. W thus would have 12.5 billion parameters which leads to very high memory consumption and high computation times.

We thus need a method that projects the outer product to a lower dimensional space and also avoids computing the outer product directly. As suggested by Gao et al. (2016) for a single modality, we rely on the count sketch projection function Ψ (Charikar et al., 2002), which projects a vector $v \in \mathbb{R}^n$ to $y \in \mathbb{R}^d$. y is initialized to a vector of d zeros, and every element v_i is multiplied by its corresponding value $s_i \in \{-1, 1\}$ and added to the jth element in y, where h maps indices i in v to indices j in y. The vectors h and s are initialized randomly from uniform distribution, but remain fixed.

This allows us to project the outer product to a lower dimensional space, which reduces the number of parameters in W. To not compute the outer product explicitly, Pham and Pagh (2013) showed that the count sketch of the outer product of two vectors can be expressed as convolution of each count sketch: $\Psi(x \otimes q, h, s) = \Psi(x, h, s) * \Psi(q, h, s)$, where $*$ is the convolution operator. Additionally, the convolution theorem states that convolution in the time domain is equivalent to element-wise multiplication in the frequency domain. The convolution $x' * q'$ can be rewritten as $\text{FFT}^{-1}(\text{FFT}(x') \odot \text{FFT}(q'))$, where $\odot$

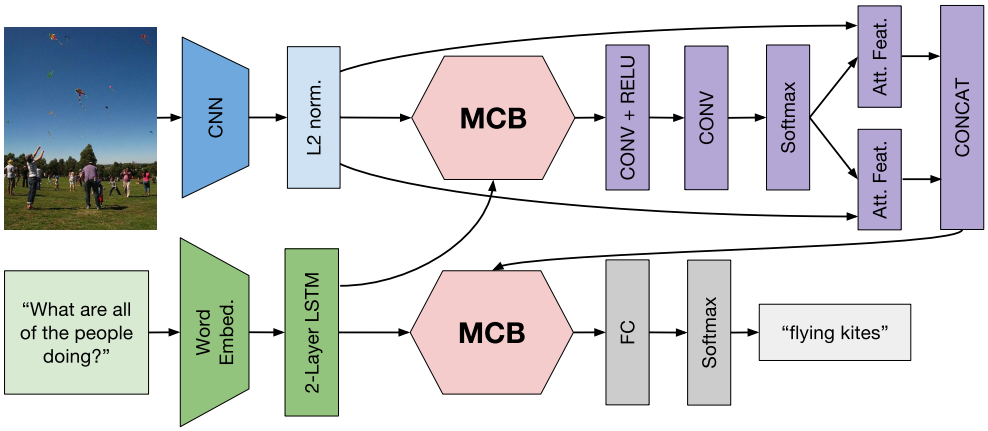

Figure 3: Our architecture for VQA: Multimodal Compact Bilinear (MCB) with Attention. CONV implies convolutional layers and FC implies fully connected layers. For details see Sec. 3.2.

refers to element-wise multiplication. These ideas are summarized in Figure 2 and formalized in Algorithm 1, which is based on the Tensor Sketch algorithm of Pham and Pagh (2013). We invoke the algorithm with $v_1 = x$ and $v_2 = q$.

3.2 Architectures for VQA

The input to the model is an image and a question, and the goal is to answer the question. Our model extracts representations for the image and the question, pools the vectors using MCB, and arrives at the answer by treating the problem as a multi-class classification problem with 3,000 possible classes.

We extract image features using a 152-layer Residual Network (He et al., 2015) that is pretrained on ImageNet data (Deng et al., 2009). Images are resized to 448×448, and we use the output of the layer ("pool5") before the 1000-way classifier. We then perform L_2 normalization on the vector.

Input questions are first tokenized into words, and the words are one-hot encoded and passed through an embedding layer. The tanh nonlinearity is used after the embedding. The embedding layer is followed by a 2-layer LSTM with 1024 units in each layer. The outputs of each LSTM layer are concatenated to form a 2048-D vector.

The two vectors are then passed through MCB. The MCB is followed by an element-wise signed square-root and L_2 normalization. After MCB pooling, a fully connected layer connects the resulting 16,000-D multimodal representation to the 3,000 top answers.

Attention. To incorporate spatial information, we use soft attention on our MCB pooling method. Explored by (Xu et al., 2015) for image captioning and by (Xu and Saenko, 2015) and (Yang et al., 2015) for VQA, the soft attention mechanism can be easily integrated in our model.

For each spatial grid location in the visual representation (i.e. last convolutional layer of ResNet [res5c], last convolutional layer of VGG [conv5]), we use MCB pooling to merge it with the language representation. As depicted in Figure 3, after the pooling we use two convolutional layers to predict the attention weight for each grid location. We apply softmax to produce a normalized soft attention map. We then take a weighted sum of the spatial vectors using the attention map to create the attended visual representation. As we can see from Figure 3, we experiment with generating two attention maps to allow the model to make multiple "glimpses" and these glimpses are concatenated before being merged with the language representation through another MCB pooling for prediction. Predicting attention maps with MCB pooling allows the model to effectively learn how to attend to salient locations based on both visual and language representations.

Answer Encoding. For VQA with multiple choices, we can additionally embed the answers. For that we base our approach on the proposed MCB with attention. As can be seen from Figure 4, to deal with multiple variable-length answer choices written in natural language, we devise an architecture in which each candidate is encoded using word embedding

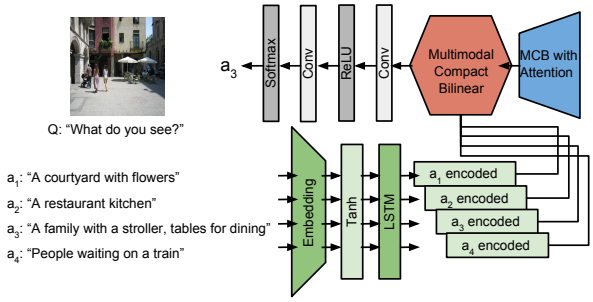

Figure 4: Our architecture for VQA: MCB with Attention and Answer Encoding

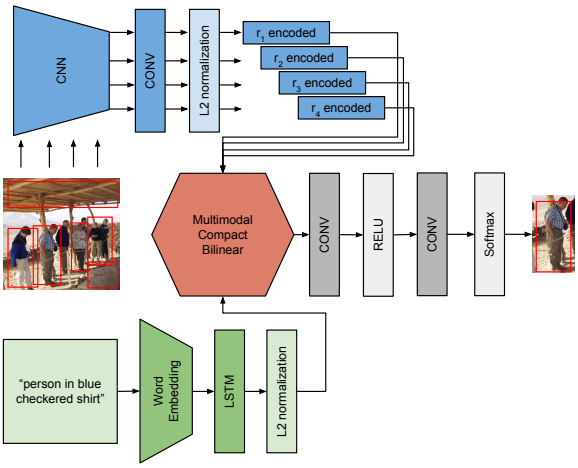

Figure 5: Our Architecture for Grounding with MCB (Sec. 3.3)

and LSTM layers whose weights are shared across the candidates. Unlike the MCB with attention, we use an additional MCB pooling to merge the encoded answer choices and the multimodal representation of the original pipeline. The resulting embedding is projected to a classification vector with a dimension equal to the number of answers.

3.3 Architecture for Visual Grounding

We base our grounding approach on the fully-supervised version of GroundeR (Rohrbach et al., 2016). The overview of our model is shown in Figure 5. The input to the model is a query natural language phrase and an image along with multiple proposal bounding boxes. The goal is to predict a bounding box which corresponds to the query phrase. We replace the concatenation of the visual representation and the encoded phrase in Grounder with MCB to combine both modalities. In contrast to Rohrbach et al. (2016), we include a linear embedding of the visual representation and L_2 normalization of both input modalities, instead of batch normalization (Ioffe and Szegedy, 2015), which we found to be beneficial when using MCB for the grounding task.

4 Evaluation on Visual Question Answering

We evaluate the benefit of MCB with a diverse set of ablations on two visual question answering datasets.

4.1 Datasets

The **Visual Question Answering** real-image dataset (Antol et al., 2015) consists of approximately 200,000 MSCOCO images (Lin et al., 2014), with 3 questions per image and 10 answers per question. There are 3 data splits: train (80K images), validation (40K images), and test (80K images). Additionally, there is a 25% subset of test named test-dev. Accuracies for ablation experiments in this paper are reported on the test-dev data split. We use the VQA tool provided by Antol et al. (2015) for evaluation. We conducted most of our experiments on the open-ended real-image task. In Table 5, we also report our multiple-choice real-image scores.

The **Visual Genome** dataset (Krishna et al., 2016) uses 108,249 images from the intersection of YFCC100M (Thomee et al., 2015) and MSCOCO. For each image, an average of 17 question-answer pairs are collected. There are 1.7 million QA pairs of the 6W question types (*what, where, when, who, why*, and *how*). Compared to the VQA dataset, Visual Genome represents a more balanced distribution of the 6W question types. Moreover, the average question and answer lengths for Visual Genome are larger than VQA. To leverage the large amount of QA pairs in the Visual Genome, we remove all the unnecessary words such as "a", "the", and "it is" from the answers to decrease the length of the answers and extract QA pairs whose answers are single-worded. The extracted data is filtered again based on the answer vocabulary space created from the VQA dataset.

The **Visual7W** dataset (Zhu et al., 2016) is a part of the Visual Genome. Visual7W adds a 7th *which* question category to accommodate visual answers, but we only evaluate the models on the Telling task which involves 6W questions. The natural language answers in Visual7W are in a multiple-choice format

Method	Accuracy
Eltwise Sum	56.50
Concat	57.49
Concat + FC	58.40
Concat + FC + FC	57.10
Eltwise Product	58.57
Eltwise Product + FC	56.44
Eltwise Product + FC + FC	57.88
MCB ($2048 \times 2048 \rightarrow 16$K)	**59.83**
Full Bilinear ($128 \times 128 \rightarrow 16$K)	58.46
MCB ($128 \times 128 \rightarrow 4$K)	58.69
Eltwise Product with VGG-19	55.97
MCB ($d = 16$K) with VGG-19	**57.05**
Concat + FC with Attention	58.36
MCB ($d = 16$K) with Attention	**62.50**

Table 1: Comparison of multimodal pooling methods. Models are trained on the VQA train split and tested on test-dev.

Compact Bilinear d	Accuracy
1024	58.38
2048	58.80
4096	59.42
8192	59.69
16000	**59.83**
32000	59.71

Table 2: Accuracies for different values of d, the dimension of the compact bilinear feature. Models are trained on the VQA train split and tested on test-dev. Details in Sec. 4.3.

No. of attention maps	Accuracy
1	64.67
2	**65.08**
4	64.24

Table 3: Accuracies for models with different numbers of attention maps. Models are trained on the VQA train, VQA validation, and Visual Genome splits and tested on test-dev. Details in Sec. 4.3.

and each question comes with four answer candidates, with only one being the correct answer. Visual7W is composed of 47,300 images from MSCOCO and there are a total of 139,868 QA pairs.

4.2 Experimental Setup

We use the Adam solver with $\epsilon = 0.0007$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use dropout after the LSTM layers and in fully connected layers. For all experiments, we train on the VQA train split, validate on the VQA validation split, and report results on the VQA test-dev split. We use early stopping: if the validation score does not improve for 50,000 iterations, we stop training and evaluate the best iteration on test-dev.

For the Visual7W task, we use the same hyper-parameters and train settings as in the VQA experiments. We use the splits from (Zhu et al., 2016) to train, validate, and test our models.

For VQA multiple choice, we trained open-ended models and took the argmax over the multiple choice answers at test time. For Visual7W, we used the answer encoding as described in Sec. 3.2.

The accuracy was computed using the evaluation code released by (Zhu et al., 2016).

4.3 Ablation Results

We compare the performance of non-bilinear and bilinear pooling methods in Table 1. We see that MCB pooling outperforms all non-bilinear pooling methods, such as eltwise sum, concatenation, and eltwise product.

One could argue that the compact bilinear method simply has more parameters than the non-bilinear pooling methods, which contributes to its performance. We compensated for this by stacking fully connected layers (with 4096 units per layer, ReLU activation, and dropout) after the non-bilinear pooling methods to increase their number of parameters. However, even with similar parameter budgets, non-bilinear methods could not achieve the same accuracy as the MCB method. For example, the "Concat + FC + FC" pooling method has approximately $4096^2 + 4096^2 + 4096 \times 3000 \approx 46$ million parameters, which matches the 48 million parameters available in MCB with $d = 16000$. However, the performance of the "Concat + FC + FC" method is only 57.10% compared to MCB's 59.83%.

Section 2 in Table 1 also shows that compact bilinear pooling has no impact on accuracy compared to full bilinear pooling. Section 3 in Table 1 demonstrates that the multimodal compact bilinear layer brings improvements regardless of the image CNN used. We primarily use ResNet-152 in this paper, but MCB also improves performance if VGG-19 is used. Section 4 in Table 1 shows that our soft attention model works best with MCB pooling. In fact,

	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.7	36.9	49.0	61.1	-	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	-	-	-	-	-	-
MCB + Genome + Att.	81.7	38.2	57.0	65.1	-	-	-	-	-	-
MCB + Genome + Att. + GloVe	82.3	37.2	57.4	65.4	-	-	-	-	-	-
Ensemble of 7 Att. models	**83.4**	**39.8**	**58.5**	**66.7**	**70.2**	**83.2**	**39.5**	**58.0**	**66.5**	**70.1**
Naver Labs (2nd best on server)	83.5	39.8	54.8	64.9	69.4	83.3	38.7	54.6	64.8	69.3
HieCoAtt (Lu et al., 2016)	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
DMN+ (Xiong et al., 2016)	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
FDA (Ilievski et al., 2016)	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
D-NMN (Andreas et al., 2016a)	81.1	38.6	45.5	59.4	-	-	-	-	59.4	-
AMA (Wu et al., 2016)	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
SAN (Yang et al., 2015)	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
NMN (Andreas et al., 2016b)	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
AYN (Malinowski et al., 2016)	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
SMem (Xu and Saenko, 2015)	80.9	37.3	43.1	58.0	-	80.9	37.5	43.5	58.2	-
VQA team (Antol et al., 2015)	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
DPPnet (Noh et al., 2015)	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
iBOWIMG (Zhou et al., 2015)	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	62.0

Table 5: Open-ended and multiple-choice (MC) results on VQA test set compared with state-of-the-art: accuracy in %. See Sec. 4.4.

Method	What	Where	When	Who	Why	How	Avg
Zhu et al.	51.5	57.0	75.0	59.5	55.5	49.8	54.3
Concat+Att.	47.8	56.9	74.1	62.3	52.7	**51.2**	52.8
MCB+Att.	**60.3**	**70.4**	**79.5**	**69.2**	**58.2**	51.1	**62.2**

Table 4: Multiple-choice QA tasks accuracy (%) on Visual7W test set.

attending to the Concat + FC layer has the same performance as not using attention at all, while attending the MCB layer improves performance by 2.67 points.

Table 2 compares different values of d, the output dimensionality of the multimodal compact bilinear feature. Approximating the bilinear feature with a 16,000-D vector yielded the highest accuracy.

Table 3 compares models with different numbers of attention maps. We find that models with two attention maps perform the best. Visual inspection of the attention maps generated leads us to believe that an ensembling or smoothing effect occurs when using multiple maps. Example attention maps are in Figure 7.

Table 4 presents results for the Visual7W multiple-choice QA task. The MCB with attention model outperforms the previous state-of-the-art by 7.9 points overall and performs better in every category. Finally, Figure 6 shows some example responses to questions by the eltwise product model and the MCB model.

4.4 Comparison to State-of-the-Art

Table 5 compares our approach with the state-of-the-art. Our best single model uses MCB pooling with attention. Additionally, we augmented our training data with images and QA pairs from the Visual Genome dataset. Also, for this model, we concatenated the learned word embedding with pretrained GloVe vectors (Pennington et al., 2014).

Our ensemble of 7 models is 1.8 points above the next best approach on the VQA open-ended task and 0.8 points above the next best approach on the multiple-choice task (on Test-dev). Even without ensembles, our "MCB + Genome + Att. + GloVe" model still outperforms the next best result by 0.5

Method	Accuracy, %
Plummer et al. (2015)	25.30
Hu et al. (2016b)	27.80
Plummer et al. (2016)[1]	43.84
Wang et al. (2016)	43.89
Rohrbach et al. (2016)	47.70
Concat	46.50
Eltwise Product	47.41
Eltwise Product + Conv	47.86
MCB	**48.69**

Table 6: Grounding accuracy on Flickr30k Entities dataset.

Method	Accuracy, %
Hu et al. (2016b)	17.93
Rohrbach et al. (2016)	26.93
Concat	25.48
Eltwise Product	27.80
Eltwise Product + Conv	27.98
MCB	**28.91**

Table 7: Grounding accuracy on ReferItGame dataset.

points, with an accuracy of 65.4% versus 64.9% on the open-ended task (on Test-dev).

5 Evaluation on Visual Grounding

5.1 Datasets

We evaluate our visual grounding approach on two challenging datasets. The first is Flickr30k Entities (Plummer et al., 2015) which consists of 31K images from Flickr30k dataset (Hodosh et al., 2014) with 244K phrases localized with bounding boxes. We follow the experimental setup of prior work, e.g. we use the same Selective Search (Uijlings et al., 2013) object proposals and the Fast R-CNN (Girshick, 2015) fine-tuned VGG16 features (Simonyan and Zisserman, 2014), as Rohrbach et al. (2016). The second dataset is ReferItGame (Kazemzadeh et al., 2014), which contains 20K images from IAPR TC-12 dataset (Grubinger et al., 2006) with segmented regions from SAIAPR-12 dataset (Escalante et al., 2010) and 120K associated natural language referring expressions. For ReferItGame we follow the experimental setup of Hu et al. (2016b) and rely on their ground-truth bounding boxes extracted around the segmentation masks. We use the Edge Box (Zitnick and Dollár, 2014) object proposals and visual features (VGG16 combined with the spatial features, which encode bounding box relative position) from Hu et al. (2016b).

[1] Plummer et al. (2016) achieve higher accuracy of 50.89% when taking into account box size and color. We believe our approach would also benefit from such additional features.

5.2 Experimental Setup

In all experiments we use Adam solver (Kingma and Ba, 2014) with learning rate $\epsilon = 0.0001$. The embedding size used in our model is 500 both for visual and language embeddings. In the following we use $d = 2048$ in the MCB pooling, which we found to work best for the visual grounding task.

The accuracy is measured as percentage of query phrases which have been localized correctly. The phrase is localized correctly if the predicted bounding box overlaps with the ground-truth bounding box by more than 50% intersection over union (IOU).

5.3 Results

Tables 6 and 7 summarize our results in the visual grounding task. We present multiple ablations of our proposed architecture. First we replace the MCB with simple concatenation of the embedded visual feature and the embedded phrase, resulting in 46.5% on the Flickr30k Entities and 25.48% on the ReferItGame datasets. The results can be improved by replacing the concatenation with the elementwise product of both embedded features (47.41% and 27.80%). We can further slightly increase the performance by introducing additional 2048-D convolution after the elementwise product (47.86% and 27.98%). However, even with fewer parameters, our MCB pooling significantly improves over this baseline on both datasets, reaching state-of the-art accuracy of 48.69% on Flickr30k Entities and 28.91% on ReferItGame dataset. Figure 6 (right) shows an example of improved phrase localization.

6 Conclusion

We propose to rely on Multimodal Compact Bilinear Pooling (MCB) to combine visual and text representations. For visual question answering, our architecture with attention and multiple MCBs gives significant improvements on two VQA datasets compared to state-of-the-art. In the visual grounding task,

Figure 6: Left: Example answers from the Eltwise Product model (EP) and MCB model on VQA images. Right: predicted groundings from the MCB model (top) and the Eltwise Product + Conv model (bottom) on Flickr30k Entities image.

introducing MCB pooling leads to improved phrase localization accuracy, indicating better interaction between query phrase representations and visual representations of proposal bounding boxes.

Acknowledgments

We would like to thank Yang Gao and Oscar Beijbom for helpful discussions about Compact Bilinear Pooling. This work was supported by DARPA, AFRL, DoD MURI award N000141110688, NSF awards IIS-1427425 and IIS-1212798, and the Berkeley Artificial Intelligence Research (BAIR) Lab.

References

[Andreas et al.2016a] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to compose neural networks for question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[Andreas et al.2016b] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Antol et al.2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *arXiv:1505.00468*.

[Charikar et al.2002] Moses Charikar, Kevin Chen, and Martin Farach-Colton. 2002. Finding frequent items in data streams. In *Automata, languages and programming*, pages 693–703. Springer.

[Deng et al.2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[Donahue et al.2013] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[Escalante et al.2010] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

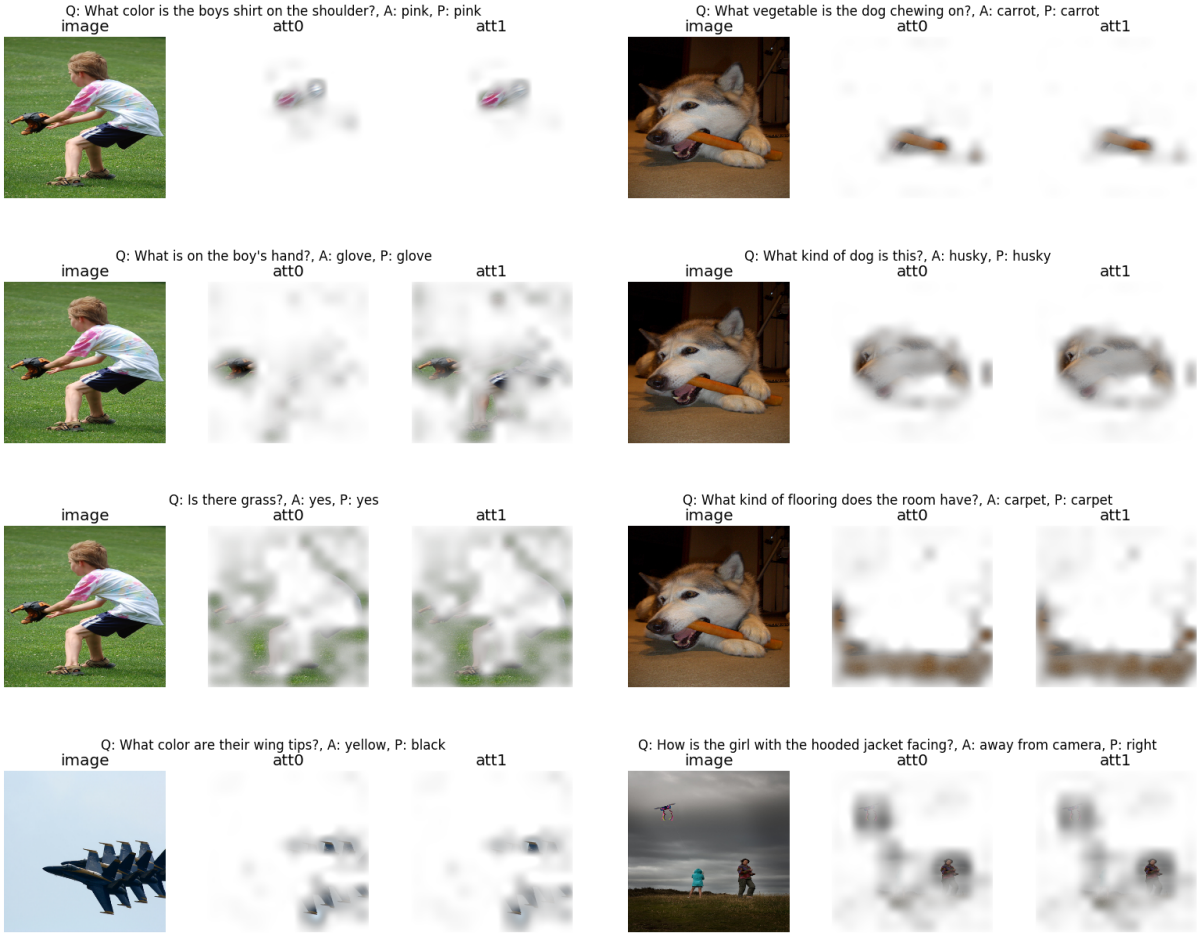

Figure 7: Visualization of attention maps for a set of questions and images from VQA validation set. "A" indicates ground-truth answer and "P" indicates model prediction. The first three rows demonstrate how the model correctly attends to different regions depending on the question. The last row demonstrates some failure cases: the first example attends to the correct region, but fails to do finegrained recognition while the second example attends to the wrong location.

[Frome et al.2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*.

[Gao et al.2016] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Girshick2015] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Gong et al.2014] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Grubinger et al.2006] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, volume 5, page 10.

[Hardoon et al.2004] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

[He et al.2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv:1512.03385*.

[Hodosh et al.2014] Peter Hodosh, Alice Young, Micah Lai, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics (TACL)*.

[Hu et al.2016a] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016a. Segmentation from natural language expressions. *arXiv:1603.06180*.

[Hu et al.2016b] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016b. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ilievski et al.2016] Ilija Ilievski, Shuicheng Yan, and Jiashi Feng. 2016. A focused dynamic attention model for visual question answering. *arXiv:1604.01485*.

[Ioffe and Szegedy2015] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*.

[Karpathy and Fei-Fei2015] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Kazemzadeh et al.2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[Kingma and Ba2014] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

[Kiros et al.2014] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 595–603.

[Kiros et al.2015] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*.

[Klein et al.2015] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Krishna et al.2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv:1602.07332*.

[Kumar et al.2015] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *arXiv:1506.07285*.

[Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Lin et al.2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Lu et al.2016] J. Lu, J. Yang, D. Batra, and D. Parikh. 2016. Hierarchical Co-Attention for Visual Question Answering. *arXiv:1606.00061*.

[Malinowski et al.2016] M. Malinowski, M. Rohrbach, and M. Fritz. 2016. Ask Your Neurons: A Deep Learning Approach to Visual Question Answering. *ArXiv e-prints*, May.

[Mao et al.2015] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of the International Conference on Learning Representations*.

[Ngiam et al.2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 689–696.

[Noh et al.2015] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2015. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv:1511.05756*.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

[Pham and Pagh2013] Ninh Pham and Rasmus Pagh. 2013. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 239–247, New York, NY, USA. ACM.

[Plummer et al.2015] Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Plummer et al.2016] Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *arXiv:1505.04870v3*.

[Rohrbach et al.2016] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. *arXiv preprint arXiv:1511.03745*.

[Simonyan and Zisserman2014] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

[Socher et al.2014] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[Tenenbaum and Freeman2000] Joshua B Tenenbaum and William T Freeman. 2000. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283.

[Thomee et al.2015] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817.

[Uijlings et al.2013] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2).

[Wang et al.2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Weston et al.2011] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770.

[Wu et al.2016] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2016. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*.

[Xiong et al.2016] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. *arXiv:1603.01417*.

[Xu and Saenko2015] Huijuan Xu and Kate Saenko. 2015. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv:1511.05234*.

[Xu et al.2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning (ICML)*.

[Yang et al.2015] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2015. Stacked attention networks for image question answering. *arXiv:1511.02274*.

[Zhou et al.2015] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv:1512.02167*.

[Zhu et al.2016] 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 (CVPR)*.

[Zitnick and Dollár2014] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–405. Springer.