Comparing Multimodal Representations in Co-Attention-Based Models for Visual Question Answering

Laura Kopf **Patrick Kahardipraja**

Abstract

Multimodal vision and language tasks such as visual question answering (VQA) and visual referring expression are challenging, because they require both the semantic understanding of image content and natural language. To solve multimodal problems, it is crucial to represent information from multiple modalities in a meaningful fashion. In this work, we investigate the properties of joint multimodal representations derived from two closely similar architectures – a task-specific model (MCAN (Yu et al., 2019)) and a multi-task model (multi-task ViLBERT (Lu et al., 2020)) – with different training objective and information streams. Our experimental results suggest that both architectures improve joint multimodal representations in an orthogonal direction. On the one hand, we show that bidirectional information streams and multi-task training in multi-task ViLBERT assist with learning underlying associations between language and visual concepts. On the other hand, we demonstrate that the attention reduction module in MCAN helps in summarizing information from multi-headed attention heads and improves conceptual groundings. Our code is publicly available at GitHub.[1]

1 Introduction

Multimodal representation of language and vision plays a key role in many language and vision tasks such as image-text retrieval (Wang et al., 2016), visual commonsense reasoning (Zellers et al., 2019), visual entailment (Xie et al., 2019), and visual question answering (Antol et al., 2015). It depicts a concept from different perspectives, which is usually complementary or supplementary in contents and therefore more informative than unimodal data. In our work, we focus on joint multimodal representations learned by visual question answering models. Visual question answering (VQA) stands out as particularly challenging compared to other language and vision tasks, as it also involves solving many subtasks like object detection, activity recognition, knowledge base reasoning, and commonsense reasoning. As such, the learned multimodal representation is rich in visual and linguistic information.

A variety of models have been developed to solve VQA, mainly utilizing various attention methods such as stacked attention networks (Yang et al., 2016), bottom-up and top-down attention mechanism (Anderson et al., 2018), and compositional attention networks (Hudson and Manning, 2018). A wave of recent work has also tried to learn textual and visual attention simultaneously through a co-attention mechanism. Such line of work (Yu et al., 2019; Nguyen and Okatani, 2018) demonstrated significantly better performance in VQA.

Recently, BERT (Devlin et al., 2019) has achieved state-of-the-art results on a multitude of NLP tasks. This leads to substantial interest in general architectures to learn joint representations of language and visual content, which can be fine-tuned on downstream tasks (ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2020), LXMERT (Tan and Bansal, 2019)). However, the need to fine-tune on downstream

[1] https://github.com/lkopf/joint-multimodal-embeddings

tasks still results in a collection of independent task-specific models instead of a single universal model. Lu et al. (2020) extended ViLBERT further with multi-task training on 12 different datasets, which results in a strong performance across a diverse set of language and vision tasks, even without further fine-tuning.

In our work, we aim to investigate the joint multimodal representation by comparing MCAN, which is a task-specific model, against multi-task ViLBERT to examine the characteristics of the learned representations. The architectures of MCAN and multi-task ViLBERT are closely similar, with the main difference being that MCAN only allows a single modality to guide another modality, while multi-task ViLBERT allows both modalities to exchange information simultaneously. Another difference lies in the multimodal fusion method, as MCAN learns to project textual and image representations to a shared space and sum them. Multi-task ViLBERT on the other hand applies the element-wise product between visual and linguistic representations. We intend to find out how these differences affect the properties of the learned multimodal representations. Additionally, we compare the grounding capabilities of both models to investigate to what extent they ground their reasoning in the image, as opposed to learning superficial correlations in the training data. In order to answer these questions, we apply BERT as embeddings and as an encoder to MCAN, replacing GloVe (Pennington et al., 2014) (with LSTM encoder) and bring it on a similar footing to multi-task ViLBERT with respect to language representation and train them on VQA 2.0 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019). For the grounding comparison, we use the grounding metric available on GQA.

Our experimental results show that the representation learned by multi-task ViLBERT is better at associating the semantic relationship between language and visual concepts. This is demonstrated by its stellar performance on VQA 2.0 and GQA. However, MCAN outperforms multi-task ViLBERT on the GQA grounding metric. This finding suggests that the attention reduction module in MCAN, which does not exist in multi-task ViLBERT, assists the model to learn better grounding capabilities by summarizing the information in multi-headed attention, and refocuses the model to the most salient information for visual question answering. Furthermore, it suggests that the learned representation by ViLBERT is weakly grounded on perception. We also show that using BERT only offers a marginal performance increase over GloVe in general.

2 Related Work

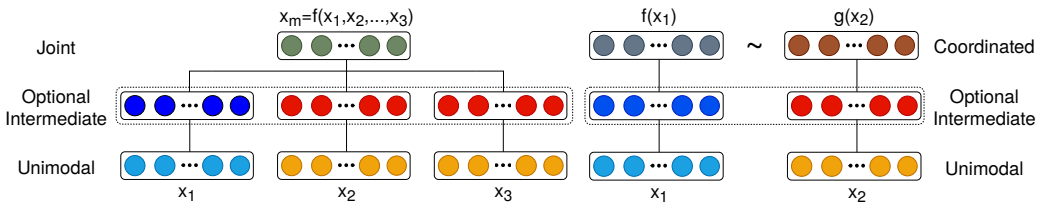
2.1 Advances in Multimodal Embeddings

Modality refers to the way in which something happens or is experienced. A research problem or dataset is characterized as multimodal when it includes multiple modalities such as language (written or spoken), visual (images, videos), or speech (sounds and para-verbal expressions). The aim of multimodal machine learning is to build models that can process and relate information from multiple modalities. The flow of multimodal information is different depending on the multimodal tasks and model architecture. Multimodal machine learning is a multidisciplinary field with a wide range of application areas such as speech recognition, event detection, emotion and affect recognition, media description, multimedia retrieval, and multimedia generation. These applications are faced with challenges such as varying levels of noise and conflicts between modalities. In this section, we will mainly focus on the challenges of multimodal representation and multimodal fusion and discuss their advancements.

2.1.1 Multimodal Representations

The challenge of multimodal representation is to learn how to represent and summarize multimodal data in a meaningful way. In order for a computational model to process data, the data first has to be transformed into a format that can be easily processed. The most commonly used format is a vector or tensor representation of an entity referring to a representation or feature. This entity can be an image, audio sample, individual word, or a sentence. There are many challenges that come with representing multiple modalities: combining data from heterogeneous sources, handling missing data, and dealing with different levels of noise. Having good representations is crucial for the performance of multimodal machine learning systems. Some properties for good performance are smoothness, sparsity, temporal and spatial coherence, and natural clustering among others (Bengio et al., 2013). Most unimodal representations nowadays are data-driven, meaning that they are learned from data using neural architectures and not hand-designed for specific applications (e.g., image features

from convolutional neural networks (CNN) (Krizhevsky et al., 2012) and textual features by word embeddings (Mikolov et al., 2013b), which we will explain more in §2.2). We will now introduce two methods of combining multimodal representations: joint and coordinated representations.



(a) Joint representation. (b) Coordinated representations.

Figure 1: Structure of joint and coordinated representations.

Joint representations project unimodal representations together into a multimodal space. In Figure 1a we can see a graph that illustrates the mathematical expression:

$$x_m = f(x_1, \ldots, x_n) \tag{1}$$

for joint representation. We can see how the unimodal representations $x_1, \ldots, x_n$ are inserted into the function f that could be e.g., a deep neural network or a recurrent neural network. This results in the joint multimodal representation x_m. Joint representations are best suited in situations where all the modalities (even more than two) are present during inference. Up until recently, most joint representations were a simple concatenation of individual modality features, which is called early fusion (D'mello and Kory, 2015). More advanced methods are neural networks, graphical models and recurrent neural networks (RNN).

Neural networks are commonly used to combine visual and textual modalities (Silberer and Lapata, 2014) or audio (Mroueh et al., 2015; Ngiam et al., 2011; Wu et al., 2014). They can be trained end-to-end, to learn both representation of the data and learning a particular task. Neural network-based joint representations have the advantage of being able to be trained from unlabeled data, if the available labeled data is not sufficient for supervised learning. One disadvantage is, that they do not naturally have the ability to deal with missing data. Variations of probabilistic graphical models on the other hand have the ability to deal with missing data in a natural way. They use latent random variables to construct representations (Bengio et al., 2013) and do not need supervised data for training (Salakhutdinov and Hinton, 2009). Both discussed models are only able to represent fixed-length data, whereas RNNs and their variants are able to represent varying length sequences. RNNs have been used in tasks such as affect recognition (Chen and Jin, 2015; Nicolaou et al., 2011) and multimodal gesture recognition (Rajagopalan et al., 2016).

Coordinated representations project each modality into a separate but coordinated space. They are coordinated through a similarity or structure constraint (e.g., minimizing cosine distance (Frome et al., 2013), maximizing correlation (Andrew et al., 2013), and enforcing a partial order (Vendrov et al., 2016) between the resulting spaces). In Figure 1b we can see the graphical illustration of the mathematical expression:

$$f(x_1) \sim g(x_2) \tag{2}$$

for coordinated representations. Here we can see that each modality (x_1, x_2) has a corresponding projection function (f, g) which is independently mapped into a coordinated multimodal space, indicated as $\sim$ in the graph. The amount of modalities has been mostly limited to two for coordinated representations. They are suited for applications where only one modality is present at inference time.

An early example for coordinated representations is WSABIE (web scale annotation by image embedding) (Weston et al., 2011), where similarity was enforced between image representations and their annotations through a coordinated space. This was attained through higher inner product, which reduced the cosine distance between the corresponding representations. A newer example for coordinated representations is DeViSE (deep visual-semantic embedding) (Frome et al., 2013), which is based on neural networks. It is similar to WSABIE, but uses more complex image and word embeddings. Another model similar to DeViSE uses videos instead of images (Pan et al., 2016).

2.1.2 Multimodal Fusion Methods

We previously described early fusion as the simplest form of joint representation. We will now discuss further methods and challenges of multimodal fusion. Multimodal fusion describes the process of joining information from two or more modalities to perform a prediction. There are several issues that might occur, e.g., that information from different modalities has varying predictive power, noise topology, and missing data in at least one of the modalities. However, multimodal fusion methods have many benefits such as allowing more robust prediction, capturing complementary information, and being able to operate, even if one modality is missing. Multimodal fusion can be classified into two main categories: model-agnostic approaches and model-based approaches. The former approach is not directly dependent on a specific machine learning approach and the latter is tied to their construction. In model-agnostic approaches there are different levels of fusion: early fusion (Leong and Mihalcea, 2011; Bruni et al., 2011), late fusion (Gunes and Piccardi, 2005; Snoek et al., 2005), and hybrid fusion (Atrey et al., 2010).

Early fusion or feature level fusion methods create a joint representation of input features from multiple modalities. Once the information is fused, a single model is trained to learn the correlation and interactions between low level features of each modality. The features are commonly concatenated, which is the simplest form of joint representation. The final prediction p is denoted as:

$$p = h([v_1, \ldots, v_m]) \tag{3}$$

where h refers to the single model and $v_1, \ldots, v_m$ represent each input modality as a dense vector. The features from different modalities need to be highly engineered and preprocessed, in order for them to align well or share similarities in their semantics. We can infer from the formula that only one model is used to make predictions, assuming that the model is well suited for all the modalities. Early fusion only requires the training of a single model, which makes the training pipeline easier compared to late and hybrid fusion.

Late fusion or decision level fusion, on the other hand, performs multimodal integration at later prediction stages. It uses unimodal decision values and fuses them with a fusion mechanism such as averaging (Shutova et al., 2016), voting schemes (Morvant et al., 2014), weighting based on channel noise (Potamianos et al., 2003) and single variance (Evangelopoulos et al., 2013), or a learned model (Glodek et al., 2011; Ramirez et al., 2011). The final prediction can be denoted as:

$$p = F(h_1(v_1), \ldots, h_m(v_m)) \tag{4}$$

where F represents a fusion mechanism and the model h_i is used on modality $i (i = 1, \ldots, M)$. The use of different models on different modalities allows for more flexibility. It makes it easier to handle a missing modality, since the predictions are made separately. However, late fusion ignores the low level interaction between the modalities and is therefore not effective at modeling signal-level interactions between modalities. In this work, we focus on late fusion with co-attention learning that models cross-modal interactions. Examples for such an approach are MCAN (Yu et al., 2019) and multi-task ViLBERT (Lu et al., 2020), which we will discuss in more detail in §2.4 and §2.5.

Hybrid fusion exploits the advantages of both feature level and decision level fusion strategies in a common framework. Its successful applications include multimodal speaker identification (Wu et al., 2005) and multimedia event detection (Lan et al., 2014).

Having presented the different fusion methods for model-agnostic approaches, we will now move on to discuss model-based approaches which can be divided into three categories: kernel-based methods, graphical models, and neural networks.

Multiple kernel learning (MKL) methods are an extension to kernel support vector machines (SVM) that are able to use different kernels for different modalities of the data (Gönen and Alpaydın, 2011). Modality-specific kernels in MKL enable better fusion of heterogeneous data, because kernel functions can be seen as similarity functions between data points. They have been an especially popular method for fusing visual descriptors for object detection (Bucak et al., 2014; Gehler and Nowozin, 2009; Krizhevsky et al., 2012) and have also been used for other tasks such as multimodal affect recognition (Chen et al., 2014; Jaques et al., 2015; Sikka et al., 2013), multimodal sentiment analysis (Poria et al., 2015), and multimedia event detection (Yeh et al., 2012). MKL are flexible in

kernel selection and can be used to both perform regression and classification. A huge disadvantage of MKL is its reliance on training data during inference time, which leads to slow inference and a large memory footprint.

Graphical models can generally be classified into two main categories: generative (modeling joint probability) and discriminative (modeling conditional probability). Over the years generative models lost popularity to discriminative ones such as conditional random fields (CRF) (Lafferty et al., 2001) and its variations (Quattoni et al., 2007; Song et al., 2012; Qin et al., 2009). Graphical models have the advantage of being able to easily exploit the spatial and temporal structure of the data. This makes them especially popular for temporal modeling tasks, such as audio-visual speech recognition (AVSR) (Gurban et al., 2008) and multimodal affect recognition (Baltrušaitis et al., 2013).

Neural networks have recently become an increasingly popular way to tackle multimodal fusion. Their field of application encompasses fusing information for visual and media question answering (Gao et al., 2015; Malinowski et al., 2015; Xu and Saenko, 2016), gesture recognition (Neverova et al., 2016), affect analysis (Kahou et al., 2015; Nojavanasghari et al., 2016), and video description generation (Jin and Liang, 2016; Venugopalan et al., 2016). Shallow (Gao et al., 2015) and deep (Nojavanasghari et al., 2016; Venugopalan et al., 2016) neural networks have both been explored for multimodal fusion, whereas the advantage of the latter lies in their capacity to learn from a large amount of data. Another advantage of recent neural architectures is their ability for end-to-end training of both the multimodal representation component and the fusion component. In comparison to other non-neural network-based systems they show good performance and are able to learn complex decision boundaries. However, they also come with disadvantages such as a requiring large training datasets to be successful and a lack of interpretability.

Although multimodal learning methods have experienced great advances in recent years, there are still some challenges that remain. The heterogeneity of data is an issue, which makes it difficult to map data from one modality to another. The relationship between modalities is often open-ended or subjective, making it hard to evaluate if there is no one correct solution. Often times multimodal learning methods can be difficult to interpret, as it might remain unclear what the prediction relies on, and which modalities or features play an important role.

2.2 Visual Question Answering

The integration of vision and language tasks has significantly advanced multi-disciplinary research from fields of computer vision, natural language processing and deep learning. The underlying challenge of this integration is to find methods that are able to process and relate information from multiple modalities, such as linguistic and visual information, according to the given task. Language and vision integration tasks are very diverse, including visual description generation (Plummer et al., 2015), image-text retrieval (Wang et al., 2016), visual commonsense reasoning (Zellers et al., 2019), visual entailment (Xie et al., 2019), and visual question answering (Antol et al., 2015). In our work, we will focus on the visual question answering (VQA) task.

VQA is a challenging multi-modal task and intersects with other vision and language tasks. The goal of VQA is to learn a model to produce a natural language answer for free-form, open-ended natural language questions by reasoning about presented visual content (as shown in Figure 2). The visual input can be either images or videos. We will only focus on images as input in further discussion. In regards to VQA demanding multi-modal knowledge beyond a single domain, it has been widely accepted as an AI-complete task. Geman et al. (2015) have considered VQA as the Visual Turing Test, where human-level abilities to semantically understand visual information and answering questions are expected. Solving the VQA task efficiently can result in various potential applications. It can for instance help blind users communicate with pictures, allow users of online educational services to interact with images, summarize visual data for surveillance data analysis and lastly through image retrieval improve the search queries on online shopping sites (Manmadhan and Kovoor, 2020).

Compared to other vision and language tasks, VQA stands out as a particularly challenging task for various reasons. In order to predict an accurate answer, a VQA model requires a fine-grained semantic understanding of both the image and the question. Acquiring this involves solving a wide range of computer vision subtasks such as object recognition, object detection, attribute classification, scene classification, counting, activity recognition, spatial relationships among objects, commonsense reasoning, and knowledge-base reasoning (Manmadhan and Kovoor, 2020). Unlike other tasks,

where the question to be answered is fixed and only the image changes, the questions in VQA are not predetermined. Another challenge is the high-dimensionality of the supporting visual information.

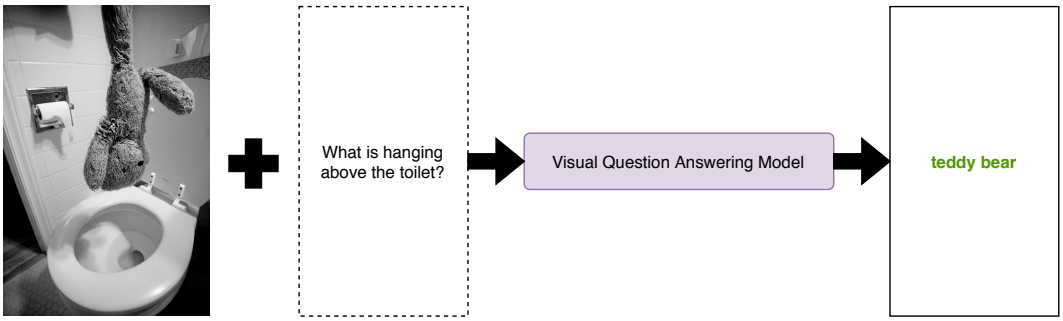


Figure 2: Given an image and question as input, the visual question answering model predicts an answer to it.

There are numerous publicly available datasets for validating VQA models with their own characteristics. The first proposed dataset for VQA is DAQUAR (DAtaset for QUestion Answering on Real-world images) (Malinowski and Fritz, 2014) which contains human question-answer pairs about images. Subsequently, several large scale datasets based on MS COCO (Common Object in Context) (Lin et al., 2014) have been proposed. COCO-QA Ren et al. (2015a) uses COCO image captions to automatically generate questions from them and produces answers of a single-word type. The VQA 1.0 dataset proposed by Antol et al. (2015) contains three questions per image and ten ground-truth answers per question. Visual Genome (Krishna et al., 2017) represents a more balanced distribution of question types and has a larger average question and answer lengths than the VQA 1.0 dataset. Visual7W (Zhu et al., 2016) is a part of Visual Genome and adds a 7th 'which' question category to accommodate visual answers. A significant deficiency in the previously mentioned VQA datasets is that they are biased. The VQA 2.0 dataset (Goyal et al., 2017) attempted to reduce language bias by asking the same question for two images and instructing human annotators to give opposite answers. We will discuss the VQA 2.0 dataset in more detail in §4.1.

Agrawal et al. (2018) argued that VQA systems are largely driven by superficial correlations in the training data and lack sufficient visual grounding. They found that these VQA systems highly rely on language priors which encourage the systems to blindly output the most common answers by only focusing on the questions without reasoning about the visual content. For example, since about 40% of questions that begin with "what sport" have the answer "tennis", systems tend to learn to output "tennis" for these questions regardless of image content (Wu and Mooney, 2019). In order to counter the language priors in VQA datasets, Agrawal et al. (2018) proposed the VQA-CP dataset where they reorganize the training and validation splits of the VQA 1.0 and VQA 2.0 datasets in a way that the distribution of the question-answer pairs is different to the test set. Kafle and Kanan (2017) developed the dataset TDIUC (Task Directed Image Understanding Challenge) to avoid some of the limitations of previous VQA datasets such as unbalanced question types, questions that can be answered by ignoring images, and difficult evaluation process. Their proposed dataset includes more balanced questions and introduces absurd questions forcing a VQA system to determine if a question is valid for a given image. In addition to that, they introduced new evaluation metrics to compensate for biases in VQA datasets. However, the most used evaluation metric for state-of-the-art VQA models is accuracy, which represents the ratio of the number of correctly answered questions to the number of total questions.

Within the last couple years many researchers have proposed different solutions for the VQA task that commonly follow the same structure. The general VQA algorithm can be divided up into three phases: Firstly image featurization and question featurization, secondly joint comprehension, and lastly answer generation. In the first phase, the given image and question are processed independently to obtain separate vector representations. There are multiple ways to extract information about images and questions of which we will only present a selected few.

2.2.1 Image and Question Featurization

In the process of image featurization the system needs to extract relevant features of the image to understand the image content. The image feature describes an image as a numerical vector, in order for it to be applied to different mathematical operations. Most VQA models use pre-trained deep neural network models for image featurization, of which convolutional neural network (CNN) (Krizhevsky et al., 2012) pre-trained on ImageNet (Russakovsky et al., 2015) is the most widely used one due to its good performance. The predominant CNN models trained on ImageNet include AlexNet (Krizhevsky et al., 2012), ZFNet (Zeiler and Fergus, 2014), VGGNet (Simonyan and Zisserman, 2015), GoogleNet (Szegedy et al., 2015) and lastly ResNet (He et al., 2016).

Plain text or strings cannot be processed by most machine learning algorithms and almost all deep learning architectures. This demands questions to be prepared in a way that can be processed by the system. Word embeddings enable the necessary preprocessing for question featurization. Embeddings can be defined as numerical vectors that represent words or phrases from a vocabulary. These vectors represent a collection of features that hold information about the relation between words. Since word embeddings are trained on word co-occurrence, they capture semantic, morphological or contextual information. Different training algorithms and text corpora have an influence on the generated word embeddings. This makes it challenging to choose the best embedding for the VQA task.

Early embedding models have been count-based models such as one-hot vector and co-occurrence matrix (Miller and Charles, 1991). They are simple to implement and interpret, but quickly run into issues concerning their fast growth as the length of the sparse vector is generally the size of the vocabulary. An alternative method for representing a word is to use a short and dense vector. Prediction-based models such as CBOW and skip-gram (also called word2vec) (Mikolov et al., 2013a) directly learn word representation and use a neural network as their basic component to train a classifier to predict a target word. In the last few years hybrid models that combine count-based and prediction-based methods to produce a word embedding have become more popular in NLP research. One prominent example are global vectors (GloVe) (Pennington et al., 2014), which perform more accurately than skip-gram, because the global corpus statistics are captured directly by the model. An even newer form of embeddings are contextualized word representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019).

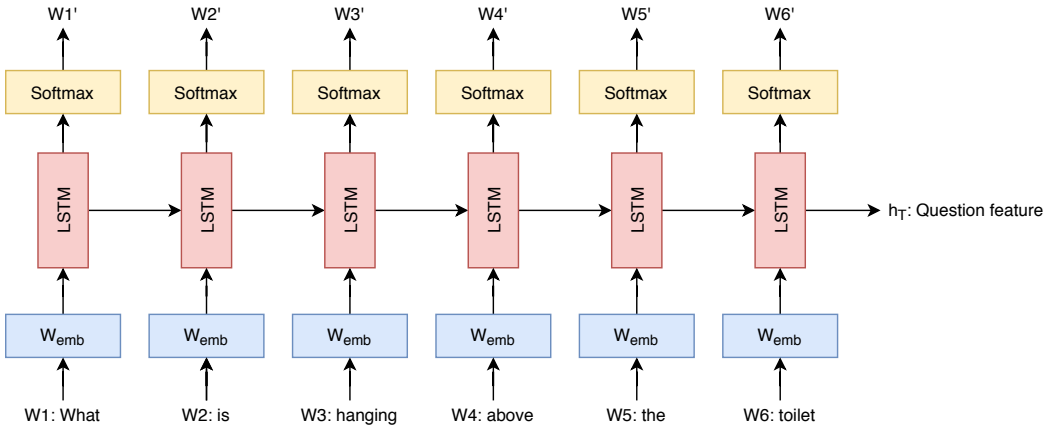


Figure 3: Basic architecture of LSTM question feature extraction.

In advanced methods for question feature extraction in VQA, neural networks such as convolutional neural network (Krizhevsky et al., 2012), long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014) are used. The latter two belong to the recurrent neural network (RNN) family. Young et al. (2018) state that sequence-based models such as RNN perform better than word sequence-independent methods like word2vec. They also claim that LSTM is generally preferred by VQA researchers. It should also be noted that RNN models are not used independently, but are always combined with any previously mentioned embeddings that are passed as input through a LSTM or GRU. Figure 3 shows the basic architecture of a LSTM where word embeddings are the input and a question feature is generated as output. Both LSTM

and GRU are gating based architectures that are designed to capture long-range dependencies and solve the vanishing gradient problem. The LSTM layer has memory cells where it can store context information e.g., of words in a sequence, when the input is a question. The memory is controlled by gates, of which one gate at each input state decides how much of the new input should be written to the memory cell, and how much of the current content of the memory cell should be forgotten. In Figure 3 h_T represents the output state vector from the last time step that is used as a question feature. GRU is an alternative to LSTM, which has fewer gates and does not have separate memory cells.

2.2.2 Joint Comprehension

After the image features and question features are extracted, the features are mapped to a joint space and then combined to generate an answer to the question about an image. There is a wide range of techniques for joint comprehension of image and text of which some were already discussed in §2.1. We will present a selection of methods for combining multi-modal features that are particularly relevant for the VQA task.

Examining several baseline fusion methods for VQA such as concatenation, element-wise addition and element-wise multiplication, Malinowski et al. (2017) have found that element-wise multiplication has more accuracy. A more advanced method for VQA is presented by end-to-end deep neural network models. They aim to capture the associations between the modalities better by training specific layers for joint comprehension of image and question features. The composition and use of the layer vary for different models. An interesting example for this method is introduced by Fukui et al. (2016), which used a Multimodal Compact Bilinear Pooling (MCB) layer for joint representation of image and question features. They hypothesize that using the outer product of visual and textual vectors is more expressive than simple baseline fusion methods. Another approach to combine multimodal features is based on encoder-decoder architecture. The decoder takes the image and question representations as input and is then trained to generate a correct answer. LSTM networks are commonly used as decoder and can vary in the way they take feature vectors as input. Ren et al. (2015a) and Zhu et al. (2016) took image encoding as first or last word of the question as input to the decoder LSTM. Malinowski et al. (2017) on the other hand took image encoding along with each word of the question.

The application of attention mechanisms has become widely popular for VQA. It enables the model to ignore image regions that are irrelevant to the given question and choose to focus on important image regions for predicting the correct answer. In the early stage attention mechanisms were mainly used for visual attention, where the focus is on regions of the image. An example of this is Stacked Attention Networks (Yang et al., 2016) that learn attention on image regions through multiple iterations.

A more advanced method for VQA is co-attention, which not only requires to learn visual attention on the image but also needs to learn textual attention on the question. One such co-attention learning method was proposed by Lu et al. (2016) that alternately learns image attention and question attention. Another co-attention model was introduced by Yu et al. (2018) where the learning method is separated into two steps. One is self-attention learning of the question, and the other is question-guided-attention learning of the image. In their approach, multiple attention maps can also be used to improve the capacity of the attended visual representation, where it is fused at a later stage with the attended question representation through multi-modal factorized high-order pooling (MFH). Nguyen and Okatani (2018) proposed the dense co-attention network (DCN) which establishes bidirectional interactions between textual and visual modalities by generating an attention map on question words for each image region and vice versa. Kim et al. (2018) introduced bilinear attention networks (BAN), where a bilinear attention map is used to reduce the computational cost to learn attention distribution for every pair of question words and image regions. They also used low-rank bilinear pooling to produce joint question and image representation. However, MFH lacks dense interaction modeling between questions and images. DCN and BAN also do not model intra-modal attention. Gao et al. (2019) and Yu et al. (2019) proposed new models based on deep co-attention that achieve a better performance on the VQA task.

2.3 Grounding

One of the main challenges of a successful VQA system is to identify and localize the most relevant image regions to the question. As discussed in §2.2.2, this is commonly resolved by attention

mechanisms. In order to find the regions of the image that lead to the answer, attention mechanisms generate an attention map over the input image. These attention maps are interpreted as groundings of the answer to the most relevant areas of the image (Zhang et al., 2019). Some questions might require external, commonsense knowledge to answer them correctly, which can make it impossible for a VQA system to ground their decision in the image that mimic human interpretation. Given an image showing a red fire hydrant and the question 'What can the red object on the ground be used for?' the system needs to visually recognize the 'red object' as a 'fire hydrant', but also to know that 'a fire hydrant can be used for fighting fires'. Another instance where VQA systems generally fail is on questions requiring reading (Singh et al., 2019). Given an image showing a ketchup bottle and the question 'What is the brand of this product?' the system needs to visually recognize the 'product' as 'ketchup bottle', but also be able to read the text on the bottle. If there is more text on the bottle, it also needs to evaluate which text is relevant for grounding the answer to the question.

To get to one of the underlying issues of this we need to take a closer look at how word embeddings get their meaning. Word embeddings are modeled through distributional semantics where the meaning of a word is entirely constituted by patterns of co-occurrence with other words or other linguistic contexts (Baroni, 2016). From a cognitive perspective, Barsalou (1999) and Fincher-Kiefer (2001) have argued that language is grounded in physical reality and perceptual experience. Barsalou (2008) has coined the term *conceptual grounding* that refers to the idea that language is grounded in perceptual experience and sensorimotor interactions with the environment.

Baroni (2016) claimed that distributional semantic models do not have access to the sensorimotor world and are therefore affected by the symbol grounding problem (Harnad, 1990). Grounding cannot be established in distributional semantics, because the meaning of linguistic symbols is given by a distribution over other linguistic symbols, which leads to infinite regress. One way to disrupt this infinite regress would be to establish an interaction with the world through perceptual and motor properties. Multimodal models that combine textual and visual modalities have the potential to establish that missing link, and allow conceptual grounding. Beinborn et al. (2018) suggested that multimodal concept representations are motivated by the idea that semantic relations between words are grounded in perception. One challenge of conceptual grounding is providing a multimodal representation of abstract concepts due to the lack of perceptual patterns associated with abstract words (Hill et al., 2014). The occurrence of abstract concepts in questions can lead to issues in the VQA task, as these might not be directly referable to the content of the image. Hence, it is crucial to combine the meaning for concrete and abstract concepts for grounding phrases.

The selection of VQA dataset plays a key role in gaining a deeper insight into how successful multimodal models are at conceptual grounding. Many VQA datasets only require relatively shallow image understanding to answer the questions which very likely enables models to get a high accuracy without requiring strong conceptual grounding abilities. Going back to our introductory examples, there are VQA datasets that specifically address issues such as reasoning about questions which require commonsense, or basic factual knowledge (Wang et al., 2018), and reading text in images (Singh et al., 2019).

2.4 Modular Co-Attention Network (MCAN)

In VQA, extracting discriminative features for textual and image representations is important in order to obtain a fine-grained semantic understanding of both the image and the question. However, using global features extracted from the whole image to represent visual information may introduce noisy information that is irrelevant to the question (e.g., the case where only a small region of the image relates to a question). On the other side, natural language questions may also contain words that are not relevant to the image and therefore can also be considered as noise. These challenges lead to the development of the co-attention learning approach, where the model jointly learns the attentions for both the image and the question simultaneously and allows it to extract more discriminative visual and textual representations.

To address the issues with previous methods for co-attention learning discussed in §2.2.2, Yu et al. (2019) proposed a modular co-attention network (MCAN), which simultaneously models dense intra- and inter-modal interactions. MCAN is composed of stacked modular co-attention (MCA) layers which consist of the self-attention (SA) and guided-attention (GA) unit based on scaled dot-product attention (Vaswani et al., 2017). The SA unit consists of a multi-head attention layer and a feed-forward neural network (FFNN) layer. It accepts a group of input features $X = [x_1, \ldots, x_m]$,

where m is the number of features, and passes it through the multi-head attention module to learn the pairwise relationship between every possible pairing $\langle x_i, x_j \rangle$ in X to obtain the attended output features Z using weighted summation of X. The output of the multi-head attention layer Z is then fed through two fully-connected layers with ReLU activation (Nair and Hinton, 2010) and dropout units (Srivastava et al., 2014). A residual connection (He et al., 2016) is employed around each of the two layers, followed by layer normalization (Ba et al., 2016) to improve optimization.

The GA unit is almost similar to the SA unit. It is composed of a multi-head attention and a FFNN layer. The GA unit accepts two groups of input features X and $Y = [y_1, \ldots, y_n]$, where n is the number of features for another modality (i.e. if X is a question embedding, Y should be an image embedding and vice versa). The GA unit learns to model the pairwise relationship between every possible pairing $\langle x_i, y_j \rangle$ from X and Y. The output Z for the GA unit can be understood as attended features for X guided by Y. In order to guide the attention learning for the GA unit key and value matrices for the multi-head attention are computed with respect to Y, while queries are computed with respect to X.

In a sense, the attended output feature $z_i \in Z$ can be seen as a reconstruction of $x_i \in X$: 1) by all $x \in X$ with respect to their normalized intra-modal similarities to x_i for the SA unit and 2) by all $y \in Y$ with respect to their normalized cross-modal similarity to x_i for the GA unit. Both SA and GA unit can be modularly combined to obtain various configurations. For instance, 2 SA units can be used to model a dense intra-modal interaction between each question word pair Y and each image region pair X separately. Afterwards, the attended visual and question features are fed to a GA unit to model dense inter-modal interactions between each word with each image region. We refer to this configuration as SA(Y) - SGA(X,Y) in the paper.

In this model, a single-layer LSTM encoder (Hochreiter and Schmidhuber, 1997) is used to compute question representations from GloVe embeddings (Pennington et al., 2014). Instead of using only the last hidden state of the LSTM encoder, MCAN utilizes the hidden states for all time steps as question embedding, yielding a feature matrix $I \in \mathbb{R}^{n \times d_i}$ where n is the number of words in the question and d_i is the size of the LSTM hidden units. For image representation, a set of regional visual features in a bottom-up manner (Anderson et al., 2018) is extracted from a Faster R-CNN (Ren et al., 2015b) with a ResNet-101 backbone (He et al., 2016). Each k-th object is represented as d_j-dimensional feature vector by mean-pooling the convolutional feature from its detected region, resulting in an image feature matrix $J \in \mathbb{R}^{m \times d_j}$, where m is the number of detected objects.

There are two variants of deep co-attention models using the MCA layer: *stacking* and *encoder-decoder*. In the *stacking* model, L MCA layers are stacked in depth with $Z_I^{(L)}$ and $Z_Q^{(L)}$ as the final attended image and question features respectively. The input features are passed recursively formulated as follows:

$$[X^{(l)}, Y^{(l)}] = \text{MCA}^{(l)}([X^{(l-1)}, Y^{(l-1)}]) \tag{5}$$

On the other side, the *encoder-decoder* model is derived from the Transformer architecture (Vaswani et al., 2017). Instead of using $Y^{(l)}$ as input features to guide the attention learning of the GA unit for each l-th layer of MCA, only the question features from the last MCA layer $Y^{(L)}$ are utilized. This can be seen as an encoder learning the attended question features $Y^{(L)}$ and a decoder learning the attended image features $X^{(L)}$ conditioned on $Y^{(L)}$. The computation for the *encoder-decoder* model is as follows:

$$Y^{(l)} = \text{SA}^{(l)}(Y^{(l-1)}) \tag{6}$$

$$X^{(l)} = \text{SGA}^{(l)}([X^{(l-1)}, Y^{(L)}]) \tag{7}$$

using SA(Y) - SGA(X,Y) configuration as an example. In both variants, input features $X^{(0)}$ and $Y^{(0)}$ are set to X and Y, respectively.

In the last phase, an attentional reduction model with a two-layer FFNN with ReLU activation and dropout units is used to obtain the final attended features $\tilde{x}$ and $\tilde{y}$ separately. Both $\tilde{x}$ and $\tilde{y}$ are

obtained as the following:

$$\alpha^x = \text{softmax}(\text{FFNN}_x(X^{(L)})) \tag{8}$$

$$\alpha^y = \text{softmax}(\text{FFNN}_y(Y^{(L)})) \tag{9}$$

$$\tilde{x} = \sum_{i=1}^{m} \alpha_i^x x_i^{(L)} \tag{10}$$

$$\tilde{y} = \sum_{i=1}^{n} \alpha_i^y y_i^{(L)} \tag{11}$$

where α^x and α^y are the learned attention weights for X and Y respectively. The multimodal fusion feature is then computed as element-wise sum:

$$z = \text{LayerNorm}(W_x^\top \tilde{x} + W_y^\top \tilde{y}) \tag{12}$$

where $W_x, W_y \in \mathbb{R}^{d \times d_z}$ are two transformation matrices and d_z is the dimension of the fused feature. The fused feature z can then be used to make predictions by projecting it to a vector $s \in \mathbb{R}^N$ followed by sigmoid activation, where N is the number of the most frequent answers obtained from the training set.

2.5 Vision-and-Language BERT (ViLBERT)

Recently, BERT (Devlin et al., 2019) has achieved state-of-the-art results on a wide array of NLP tasks, such as question answering, natural language inference, and named entity recognition, without any substantial task-specific architecture modifications. Lu et al. (2019) proposed Vision-and-Language BERT (ViLBERT), which is an extension of BERT to jointly learn task-agnostic visual grounding and reason about text and images.

In ViLBERT, two parallel BERT-style architectures are used to model intra-modal interactions and fuse them through attention-based cross-modal interactions. Besides standard transformer blocks (Vaswani et al., 2017), a novel co-attentional transformer layer is introduced to facilitate information exchange between different modalities. The co-attentional transformer layer enables simultaneous attention learning by computing key and value matrices from each modality and passing them as input to the multi-head attention layer of the other modality. As a consequence, the multi-head attention layer produces attended image features conditioned on language and attended language features conditioned on images. Both attended features are then fed through a FFNN according to their respective stream with a residual connection with their initial representations, followed by layer normalization (similar to the standard transformer layer). The model itself is composed of alternating transformer blocks and co-attentional transformer layers stacked in series.

ViLBERT takes as input an image I and text segment W where both of them are represented as the sequence $\{\texttt{IMG}, v_1, \ldots, v_\tau, \texttt{CLS}, w_1, \ldots, w_\tau, \texttt{SEP}\}$ where $v_i, 1 \leq i \leq \tau$ are a set of region features and $w_i, 1 \leq i \leq \tau$ are word tokens and the IMG, CLS, SEP tokens are special markers. Afterwards, the model outputs final embeddings for each input token $\{h_{\texttt{IMG}}, h_{v1}, \ldots, h_{v\tau}, h_{\texttt{CLS}}, h_{w1}, \ldots, h_{w\tau}, h_{\texttt{SEP}}\}$. $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ correspond to mean-pooled features that represent the entire image and text segment which can be used for downstream language and vision tasks.

For pre-training tasks, ViLBERT is trained with masked multi-modal modeling and multi-modal alignment prediction on the Conceptual Captions (Sharma et al., 2018) dataset. The masked multi-modal modeling task is derived from the masked language modeling task in BERT (Devlin et al., 2019). Both words and image region inputs are masked approximately 15% at random and the model is tasked to reconstruct the words and image region given the unmasked inputs. For masked image regions, the features are set to zero 90% of the time and unchanged 10% of the time. Masked words are replaced with a [MASK] token 80% of the time, a random token 10% of the time and unaltered 10% of the time.

In order to predict the masked values, a distribution over semantic classes for the corresponding image region is set as target as opposed to directly regressing the values. As supervision, the output distribution for the region from a pre-trained detection model for image feature extraction is used. The model is then trained to minimize the Kullback-Leibler divergence between these two distributions.

The objective in multi-modal alignment prediction is to present the model with image and text pairs and the model must predict whether, if the text describes the image. In order to do this, the element-wise product between $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ is computed as image-text representation. A linear layer is then added on top of the representation with sigmoid activation to predict if the image and text are aligned or not. For negative samples, either the image or the text is replaced with another one randomly sampled from the dataset. Similar to BERT, ViLBERT can be fine-tuned on downstream tasks by adding a FFNN layer on top of the final representations and train it in an end-to-end manner.

Lu et al. (2020) improved ViLBERT further with multi-task learning, resulting in a single unified model that can perform impressively on many language and vision tasks such as visual question answering, caption-based image retrieval, grounding referring expressions and visual entailment. Multi-task ViLBERT considers 4 groups of tasks to train on — vocab-based VQA (VQA 2.0 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and Visual Genome QA (Krishna et al., 2017)), image retrieval (MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015)), referring expressions (RefCOCO(+/g) (Kazemzadeh et al., 2014; Mao et al., 2016), Visual7W (Zhu et al., 2016), and GuessWhat (de Vries et al., 2017)), and multi-modal verification (NVLR^2 (Suhr et al., 2019) and SNLI-VE (Xie et al., 2018)). In the multi-modal verification task, given one or more images and a natural language sentence, the model should determine the truthness or predict the semantic relationship between the sentence and the image.

For each task, a task-specific layer is added on top of a ViLBERT model with shared parameters across all tasks. In order to accommodate this, a new task token $\texttt{TASK}_t$ is also added such that the input to ViLBERT is $\{\texttt{IMG}, v_1, \ldots, v_\tau, \texttt{CLS}, \texttt{TASK}_t, w_1, \ldots, w_\tau, \texttt{SEP}\}$. Vocab-based VQA is treated as multi-label classification, where the Hadamard product between $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ is computed and passed to a two-layer FFNN with sigmoid activation as follows:

$$P_v(A|I,Q) = \sigma(\text{FFNN}(h_{\texttt{IMG}} \odot h_{\texttt{CLS}})) \tag{13}$$

where A is the answer, I is the image and Q is the question. For image retrieval, the task-specific layer is trained in a 4-way multiple-choice setting against hard-negatives. An alignment score is computed between image-caption pairs formulated as:

$$\texttt{Rel}(I,Q) = W_i(h_{\texttt{IMG}} \odot h_{\texttt{CLS}}) \tag{14}$$

where $W_i \in \mathbb{R}^{d\times 1}$ is followed with softmax activation. The referring expressions task is viewed as a reranking of a set of image region proposals given the referring expression. The final representation h_{v_i} for each image region i is passed through a projection layer $W_r \in \mathbb{R}^{d\times 1}$ to learn a matching score between the proposed region and the matching score:

$$\texttt{Rel}(v_i,Q) = W_r h_{vi} \tag{15}$$

where in this case Q can be either a phrase, question or dialog depending on the datasets (RefCOCO(+/g), Visual7W, and GuessWhat).

Lastly, 2 different task-specific heads are defined for multi-modal verification due to a different task structure between NVLR^2 and SNLI-VE. In the case of NVLR^2, the model must determine the truthness of the statement given the images, which is framed as a classification problem. Given an embedding that encodes two image-statement pairs (I_0, Q) and (I_1, Q), the truthness probability is computed with a 2-layer FFNN as follows:

$$P_v(C|I_0,I_1,Q) = \text{softmax}\left(\text{FFNN}\left(\begin{bmatrix} h^0_{\texttt{IMG}} \odot h^0_{\texttt{CLS}} \\ h^1_{\texttt{IMG}} \odot h^1_{\texttt{CLS}} \end{bmatrix}\right)\right) \tag{16}$$

where $[\,]$ is concatenation. As for SNLI-VE, the input is an image premise and a text hypothesis, and the model must predict the relation between the image and the statement (entailment, neutral, contradiction). Using a linear layer, the element-wise product between $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ is then mapped to one of the three labels.

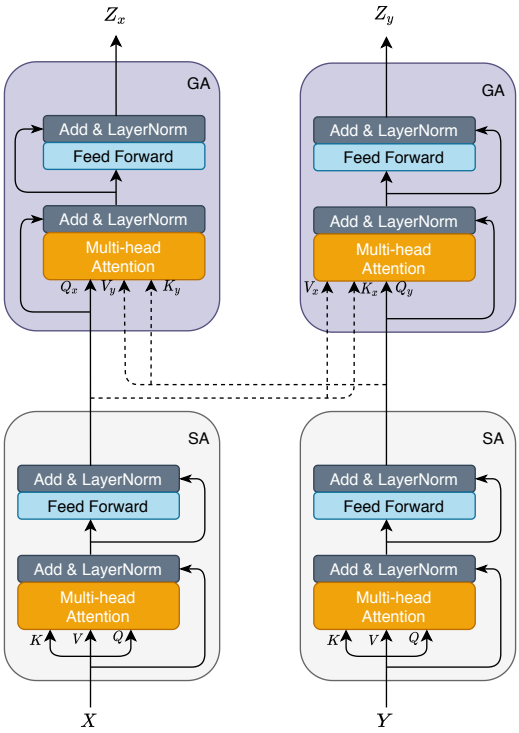
3 Approach

3.1 MCAN vs ViLBERT

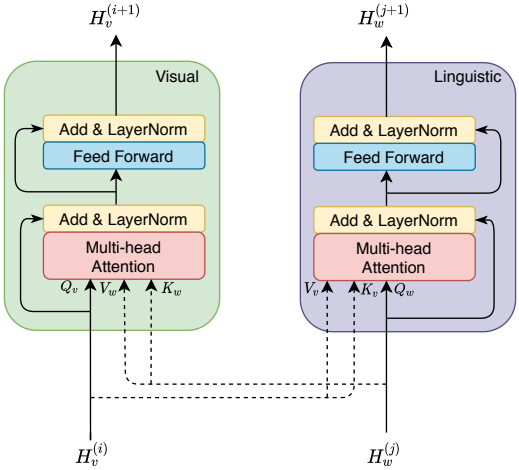
Despite being based on a transformer architecture, both MCAN and ViLBERT have several key differences. First of all, in MCAN the information exchange between modalities is unidirectional.

Only one modality can be used to guide another modality at a time (e.g., either attended image feature learning guided by question or attended question feature learning guided by image). On the other side, the novel co-attentional layer in ViLBERT allows bidirectional information exchange between language and vision at varying representation depths. Both architectures also differ in how textual and visual representations are fused. MCAN learns to project textual and image representations to a shared space and sum them together, while ViLBERT applies the element-wise product between $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ as joint representation of the visual and linguistic inputs.

Interestingly, it turns out that the co-attention mechanism used by ViLBERT can be considered as a special composition of the modular SA and GA units, as depicted in Figure 4.



(a) MCAN with SGA(X,Y) - SGA(Y,X) layer.



(b) ViLBERT co-attentional transformer layer.

Figure 4: Comparison of ViLBERT against MCAN with SGA(X,Y) - SGA(Y,X) layer. Note the similarities of information flows between the two models.

Here $H_v^{(i)}$ and $H_w^{(j)}$ are obtained from standard transformer layers. It is important to note that multi-task ViLBERT as SGA(X,Y) - SGA(Y,X) is trained with different training objectives and additional tokens which yield different visual and textual representations compared to the MCAN model.

Joint representations play an important role in visual question answering and many other language and vision tasks, since they relate important information between visual and textual modalities. As such, comparing joint representations derived from a single unified model (multi-task ViLBERT) against a task-specific model (MCAN), which share similar modular units, might reveal the importance of different training objectives and information flows in generating highly effective multimodal representations.

3.2 Applying BERT to MCAN

As opposed to ViLBERT's linguistic stream that is initialized with BERT, the original MCAN model uses GloVe as word embeddings, which is non-contextual and trained on global word-word co-occurrence statistics. However, for VQA, context also provides additional cues which may be helpful for the model to predict the answer accurately and enrich the information in multimodal embeddings (e.g., distinguishing "bank" as a financial institution and as land besides water provides extra information for the model when answering questions, especially when both appear together in a question).

Works such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) assign each word a vector as a function of the entire input sequence, which also enables them to model the use of words in a contextual manner. For our experiments, we use contextual word representations from BERT. BERT is a bidirectional language representation model trained jointly with a masked language model and next sentence prediction objective on BooksCorpus (Zhu et al., 2015) and English Wikipedia, with an approximate total of 3,300M tokens. We use BERT-base (12-layer transformers, 768-hidden) and BERT-large (24-layer transformers, 1024-hidden). This brings both MCAN and ViLBERT to an equal footing with respect to their word representations.

We apply BERT with two variants: 1) as contextualized word embeddings to replace GloVe and 2) as question encoder, by replacing the entire LSTM-based encoder (with GloVe embeddings as input) in the MCAN model with the BERT transformer. We excluded the `CLS` token when using BERT both as embedding and as encoder, as the `CLS` token is mainly included when BERT is used in a fine-tuning manner. For co-attention learning, we use the *encoder-decoder* model with SA(Y) - SGA(X,Y) layers as it performed better compared to the *stacking* strategy (Yu et al., 2019). The architecture for both variants is depicted in Figure 5.

4 Experimental Setup

We aim to investigate the characteristics of joint multimodal representations derived from a task-specific model and a single, unified model for various language and vision tasks. Furthermore, we also want to answer to what extent the models ground their reasoning in the image, rather than making guesses by memorizing questions prior from the training data. Our experiments, which we describe below, are intended to compare the joint multimodal representations and how different training objectives and information flows affect them.

4.1 Datasets

The VQA 2.0 dataset (Goyal et al., 2017) was initially developed to balance the existing VQA 1.0 dataset Antol et al. (2015) by collecting complementary images (CI) and corresponding new answers. It is almost double the size of the VQA 1.0 dataset containing approximately 1.1M question-image pairs (QI) with approximately 13M associated answers on the approximately 200K images from MS COCO (Lin et al., 2014). The question-answer pairs are annotated by humans, with three questions per image and 10 answers per question. The VQA 2.0 dataset is split into three: train (195K CI, 443K QI), val (93K CI, 214K QI), and test (191K CI, 453K QI). The test set is further split into four: *test-dev*, *test-standard*, *test-challenge* and *test-reserve*. The evaluation metric computes the accuracies for three different answer types (yes/no, number, and other) and overall accuracy.

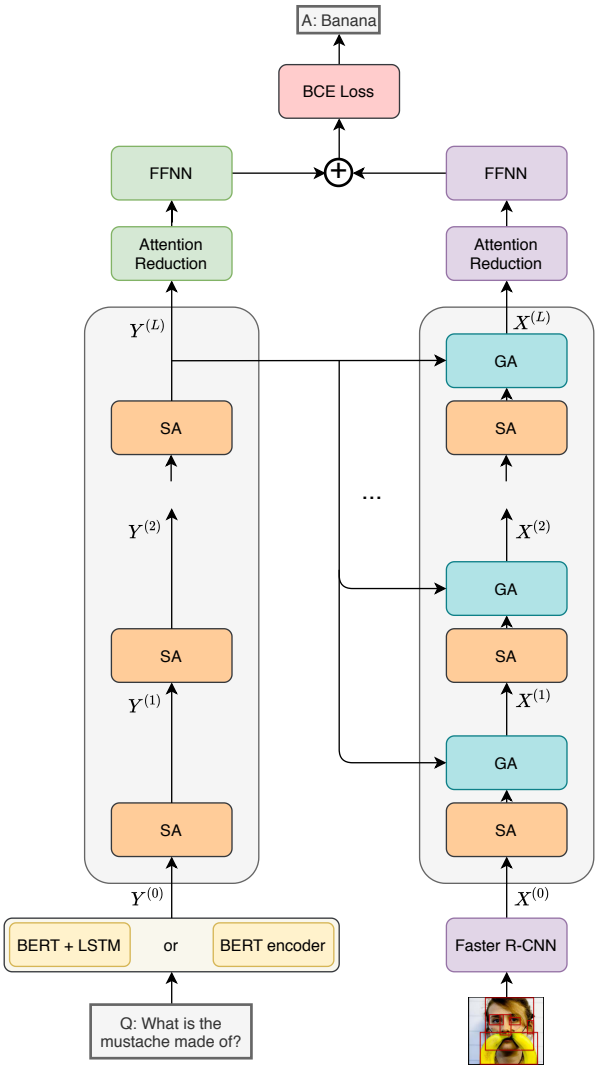


Figure 5: *Encoder-decoder* MCAN (SA(Y) - SGA(X,Y)) with BERT. We apply BERT as contextualized word representations and as question encoder, replacing a single-layer LSTM.

GQA is a dataset Hudson and Manning (2019) for visual reasoning and compositional question answering over real-world images. It consists of 113K images and 22M questions of assorted types and varying compositionality degrees. Each image is annotated with a scene graph of objects and relations. The images are from MS COCO (Lin et al., 2014) and Flickr (Thomee et al., 2016), the image scene graphs are based on Visual Genome (Krishna et al., 2017). Each question is associated with a structured representation of its semantics and has a higher average length compared to the questions in the VQA 2.0 dataset. Each answer is complemented with both textual and visual justifications, pointing to the relevant region within the image. In addition to the standard accuracy metric and the more detailed type-based diagnosis, the GQA dataset includes five new metrics. The first one being consistency, which measures the response consistency across different questions. The validity metric checks whether a given answer is in the scope of the question. The plausibility score measures whether the answer is reasonable, or makes sense, given the question. The distribution metric measures the overall match between the true answer distribution and the model predicted distribution. And lastly the grounding score, which checks whether the model attends to regions within the image that are relevant to the question.

4.2 Implementation and Hyperparameters

We extend the original MCAN implementation in OpenVQA[2] and multi-task ViLBERT[3]. We also use the HuggingFace BERT implementation (Wolf et al., 2019)[4]. When utilizing BERT as features, we freeze the weights and concatenate the four last hidden layers of BERT-base and BERT-large as opposed to the last layer, as it yields the best performance. BERT as encoder is fine-tuned in an end-to-end manner following (Devlin et al., 2019). Our MCAN model accepts input image features and input question features of size d_x = 2,048 and d_y = 512 respectively and produces fused multi-modal features with the dimensionality d_z = 1,536. Due to the dimensionality of BERT hidden state, we set d_y = 768 when applying BERT-base as encoder. We did not use BERT-large as encoder due to high computational cost. We use uncased variants of BERT in all of our experiments.

VQA image features are extracted from Faster R-CNN (with ResNet-101 backbone) while the question words are tokenized to wordpiece tokens of 14 tokens maximum following the original MCAN implementation. We use object-based features for GQA image representation, which can be downloaded from the official website. GQA questions are truncated to a maximum of 29 tokens. Due to the dynamic number of image regions m and question length n, we use zero-padding to fill the image feature matrix J and question feature matrix I, where $m_{\texttt{VQA,GQA}}$ = 100, $n_{\texttt{VQA}}$ = 14, $n_{\texttt{GQA}}$ = 29. For MCAN, the number of attention heads is set to 8 and the number of layers is set to 6. The size of the answer vocabulary is set to N = 3,129 for VQA following Yu et al. (2019) and N = 2,933 for GQA. For optimization, we use Adam (Kingma and Ba, 2014) with β_1 = 0.9 and β_2 = 0.98. The learning rate is set to $\min((lr/(w+1)) * t, lr)$ where t is the current epoch starting from 1, learning rate $lr = 7e^{-5}$ and warmup epoch w = 3 for VQA and w = 2 for GQA. We also use linear decay during training, where the learning rate is decayed by 1/5 every 2 epochs. The decay is applied for VQA after the 10th epoch and for GQA after the 8th epoch. MCAN models with BERT are trained up to 11 epochs for GQA and up to 13 epochs for VQA with batch size 64. For VQA training, we use both $train$ and val splits with additional VQA samples from Visual Genome, while for GQA, both balanced $train$ and val splits are used.

Again, due to computational constraint, we use the pre-trained 6-layer ViLBERT model trained with 12 datasets. Image features for ViLBERT are extracted from the ResNeXT-152 (Xie et al., 2016) Faster R-CNN model trained on Visual Genome with attribute loss, resulting in a maximum of 100 image regions. During VQA training, the question words are tokenized into wordpiece tokens and trimmed to a maximum of 23 tokens. For GQA, the question words are truncated to a maximum of 26 tokens. The truncation difference between MCAN and multi-task ViLBERT occurs due to the cleaned split used for multi-task training, where test images are removed from *train* or *val* split for all tasks.

4.3 Grounding

We want to understand to what extent the joint multimodal representations induced by VQA models relate visual and textual information. Do the models learn superficial correlations in the training data or do they base their reasoning on the images? In order to answer this question, we evaluate question grounding as a proxy to compare the quality of joint multimodal representations. We focus on the GQA dataset, as it offers grounding annotation in the form of a visual pointer that points to regions in the image which the question refers to and are relevant to answer it.

The question grounding metric is calculated as follows. For each question-image pair (q, i) exists one or multiple pointers r that point to relevant image regions when answering the question. The model visual attention over the image i, which is represented by object-based features in our experiments, is then intersected with pointer/s r to obtain the intersection rate. Afterwards, the intersection rate for each bounding box is multiplied with its attention score and then summed up to obtain the overall attention over r. This yields the grounding score for an instance of (q, i). The overall grounding score is then computed by averaging the grounding score over all questions in the dataset. GQA also includes spatial-based features for training and evaluation, which we did not have the opportunity to explore due to computational issues.

[2] https://github.com/MILVLG/openvqa

[3] https://github.com/facebookresearch/vilbert-multi-task

[4] https://github.com/huggingface/transformers

For MCAN, we use the attention weights obtained from the visual attentional reduction module, with $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m]$ for m bounding boxes. As for multi-task ViLBERT, we take the self-attention distribution from the last layer for the $h_{\texttt{IMG}}$ token, as it is the holistic representation of the image. This is intended to get the most accurate attention scores over the image region, as ViLBERT does not have an attention reduction module. We also examine the grounding scores for all 8 attention heads to measure how actively different heads pay attention to the image regions. We then take the average of all attention heads as final grounding score for ViLBERT. We normalize all bounding boxes for MCAN and multi-task ViLBERT.

We compare the grounding scores for MCAN against multi-task ViLBERT, as it measures the grounding capability that a task-specific model can achieve as opposed to a single model for various language and vision tasks. Furthermore, it also shows how unidirectional or bidirectional information streams for co-attention learning affect the robustness of the joint multimodal representation.

4.4 Fusion Method

While it is not possible to provide a one-to-one comparison between joint representations in MCAN and multi-task ViLBERT due to differences in architecture and learning strategy, we intend to compare both representations as fair as possible. Therefore, we focus on the cross-modal feature fusion method. MCAN learns to project textual and image representations to a shared space and then performs the element-wise sum on them. On the other hand, ViLBERT applies the element-wise product on $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ tokens to obtain joint representations. We modify the late fusion method for MCAN as the following:

$$z = \text{LayerNorm}(W_x^\top \tilde{x} \odot W_y^\top \tilde{y}) \tag{17}$$

$$\tag{18}$$

and for ViLBERT:

$$z = h_{\texttt{IMG}} + h_{\texttt{CLS}} \tag{19}$$

where z is the fused representation. Although we use the pre-trained multi-task ViLBERT that is not trained with the element-wise sum, we want to investigate the performance of the modified fusion method in visual question answering and compare it with the original fusion method.

5 Results and Discussion

5.1 Comparison on VQA 2.0

Table 1 summarizes our results for different configurations of MCAN models in comparison to multi-task ViLBERT. The baseline MCAN-small model corresponds to the one described in Yu et al. (2019) using GloVe word embeddings passed through a one-layer LSTM network with a hidden size of 512. We observe that the multi-task ViLBERT model significantly outperforms various MCAN models in all metrics.

One possible advantage of ViLBERT over our MCAN models is the bidirectional visual-linguistic stream, which enables simultaneous information exchange between different modalities. Coupled with a multi-task training objective on a plethora of datasets which are divided into 4 task groups, ViLBERT is also able to learn better associations between language and visual concepts due to parameter sharing in both base ViLBERT model and branching task-specific heads, therefore resulting in higher accuracy. Lu et al. (2020) observed that multi-task training has a regularizing effect on tasks that overfit when trained separately. This explains its benefit over the task-specific MCAN models. We find that multi-task ViLBERT is especially better compared to our best-performing MCAN model in counting by 3.42 points. This might be due to ViLBERT being trained on several VQA datasets in a round-robin fashion, which improves the model performance in counting. It is important to note that we do not apply BERT to MCAN with *stacking* SGA(X, Y) - SGA(Y, X), which is similar to the ViLBERT architecture, due to computational constraint. Yu et al. (2019) also reported that such a configuration did not offer comparative performance to SA(Y) - SGA(X, Y). Thus, we presume that the reason ViLBERT performs better is mainly due to multi-task training.

Out of all MCAN models, we observe that MCAN with BERT-base encoder performs the best in all categories, surpassing the baseline model by 0.4 points in accuracy. Applying BERT as encoder

has also achieved strong performance in tasks such as coreference resolution (Joshi et al., 2019). However, replacing GloVe with BERT as contextual word representations only yields a marginal performance increase in general, which is surprising considering how BERT significantly improves the performance in many downstream NLP tasks. The sole exception where BERT performs worse compared to GloVe is when we use BERT-large as embeddings. We attribute this to our decision in choosing the hidden state dimension of the LSTM encoder similar to the baseline model, which might be too low to model the question representation. An indication for this is that the model still does not converge, even after training it with more epochs.

We hypothesize that the performance difference when using BERT as encoder and as features is mainly due to the LSTM encoder, which is similar to the baseline model. Based on the results, we conclude that the LSTM encoder, which is unidirectional, prevents the model from making full use of the contextual information from BERT embeddings. Using a bidirectional LSTM has the potential to improve the performance further, as it might be able to incorporate the contextual information from BERT embeddings better. Another possible cause for this is that we use a single layer LSTM, which might be too shallow to properly model the question representation. Using a deeper LSTM network might be beneficial for better question modeling.

Model	Overall	Y/N	Num	Other
MCAN-small (baseline)	70.63	86.82	53.26	60.72
MCAN /w BERT-base features	70.70	86.82	52.94	60.96
MCAN /w BERT-large features	70.53	86.51	53.01	60.86
MCAN /w BERT-base encoder	71.03	87.34	53.30	61.12
Multi-task ViLBERT	72.57*	88.33*	56.72*	62.68*

Table 1: Results for MCAN models and multi-task ViLBERT on the *testdev* split of the VQA 2.0 dataset (Goyal et al., 2017). Models are evaluated for overall accuracy as well as the accuracies for three different answer types (yes/no, number, and other). Asterisk denotes the best performance on each metric.

5.2 Comparison on GQA

We compare the results for MCAN models and multi-task ViLBERT on the GQA dataset (Hudson and Manning, 2019), which is depicted in Table 2. Again, we examine that multi-task ViLBERT significantly outperforms the MCAN models in all metrics. Compared to the previous results on VQA 2.0, the gap between the performance of multi-task ViLBERT and MCAN models is substantially larger (by 5.62 points for accuracy on average).

The superior results of ViLBERT could be attributed to its bidirectional information stream and robust generalization across language and vision tasks which is achieved by multi-task training, similar to its performance on the VQA 2.0 dataset. We observe that multi-task ViLBERT achieves the best distribution score (for this metric, lower is better), which demonstrates that the model is able to predict not only the most common answers, but also the less frequent ones. This can be explained by the fact that multi-task training allows for better generalization compared to task-specific training, which in return enables ViLBERT to learn more robust representation. In comparison with the similar metric in VQA 2.0 (Y/N), multi-task ViLBERT also shows superior performance in the binary metric with a gap of 5.53 points over our best MCAN model.

Similar to the result on VQA 2.0, MCAN with BERT only offers a marginal improvement in performance across all metrics. We examine that out of all MCAN models, MCAN with BERT-base features scores the highest on overall accuracy and binary. Interestingly, MCAN with BERT-base encoder performs the worst out of our modified MCAN models, contrary to our previous VQA 2.0 results. We attribute this to our decision in using the same learning rate for MCAN and BERT encoder. We suspect this introduces catastrophic forgetting, which damages the learned language representation in BERT due to overly aggressive fine-tuning. We also did not observe this phenomenon when training MCAN with BERT encoder on VQA 2.0, which might be due to the fact that VQA 2.0 is less challenging and therefore did not cause the BERT base network to drift as much compared to when we train the model on GQA.

In comparison to the overall accuracy scores on VQA 2.0, we can see a significant drop on GQA. This is because the questions in GQA tend to be more challenging, as they are more compositional compared to VQA 2.0 and involve a wide array of reasoning skills (i.e. object and attribute recognition, transitive relation tracking, spatial reasoning, logical inference and comparisons). Hudson and Manning (2019) also argued that open questions are not balanced in the VQA 2.0 dataset. To address this issue, the open questions in GQA are balanced by applying a smoothing technique in order to make the answer distribution for each question group more uniform. As a result, GQA is more robust against shortcuts and guesses, which are often exploited by VQA models when facing difficult learning problems (Agrawal et al., 2018). As seen from Table 2, both ViLBERT and MCAN score relatively low for the open questions category. This suggests that there is still room for improvement in modeling fine-grained recognition and commonsense reasoning, which are both critical for answering open-ended questions.

Model	Accuracy	Binary	Open	Distribution
MCAN-small (baseline)	53.41	70.29	38.56	1.40
MCAN /w BERT-base features	54.79	72.96	39.37	1.70
MCAN /w BERT-large features	54.70	72.18	39.87	1.74
MCAN /w BERT-base encoder	53.62	71.12	38.77	1.77
Multi-task ViLBERT	59.75*	78.49*	43.85*	1.31*

Table 2: Results for MCAN models and multi-task ViLBERT on the GQA dataset (Hudson and Manning, 2019). All results refer to the balanced *testdev* set. Models are evaluated for overall accuracy as well as accuracies for two different answer types (binary, open). In addition, they are evaluated by the distribution metric which is calculated using Chi-Square statistic (lower is better). Asterisk denotes the best performance on each metric.

5.3 Question Grounding

Table 3 summarizes our results for question grounding on the GQA dataset. The grounding score indicates whether the model actively attends to regions in the image that are relevant for answering the question. Hence, it shows to what extent the model grounds its reasoning in the image. The MCAN model with BERT-base features performs best on grounding and significantly outperforms multi-task ViLBERT by over 30 points. This is especially interesting, as multi-task ViLBERT has consistently outperformed all MCAN models on the overall accuracy scores on VQA 2.0 and GQA. Even if we consider the best attention head of ViLBERT, there is still a difference of over 10 points to our best-performing MCAN model.

We hypothesize that the attention reduction module of MCAN helps to summarize the information in all the attention heads and refocus the model to attend to the relevant image regions. This can be examined from the grounding score of MCAN with BERT as encoder, which performs the worst but still managed to score 4.17 points higher compared to the best attention head of ViLBERT. This also indicates that joint representations learned by the task-specific model are better at conceptual grounding. Furthermore, this demonstrates that the task-specific model relies less on memorizing statistical biases in the dataset and grounds most of its reasoning in the image.

We also observe the grounding scores for all 8 attention heads of multi-task ViLBERT in order to obtain a better insight on the multi-headed attention mechanism. We find that out of all heads, only 2 heads are actively paying attention to the relevant visual regions (head 1 and 7). This is consistent with previous findings from Li et al. (2020), where they demonstrate that only certain attention heads of ViLBERT-like architectures actively ground elements of language to image regions, as the attention heads specialize in different things. They also discover that attention heads in upper layers tend to have higher grounding accuracy.

Multi-task ViLBERT might be good at generalizing due to its bidirectional information streams and training objective which has regularizing effects, but it lacks the ability to focus its attention on relevant regions of the image when answering the question. It is remarkable that multi-task ViLBERT still manages to score a high overall accuracy on GQA and VQA 2.0 compared to MCAN, despite not truly grounding its reasoning in the images. Multi-task ViLBERT's over-reliance on language

might also be influenced by its weight initialization, which is obtained by pre-training with masked multi-modal modeling and multi-modal alignment prediction. During pre-training, visual regions with significant overlap are aggressively masked to avoid leaking visual information. It seems that this, combined with the already available language representation from the ViLBERT linguistic stream, forces the model to rely more heavily on language.

Model	Attention Head	Grounding
MCAN-small (baseline)		86.34
MCAN /w BERT-base features		90.90*
MCAN /w BERT-large features		89.61
MCAN /w BERT-base encoder		83.90
Multi-task ViLBERT	1	73.35
	2	54.33
	3	48.46
	4	49.43
	5	47.52
	6	54.12
	7	79.73
	8	57.75
Multi-task ViLBERT (averaged)		58.09

Table 3: Results for MCAN models and multi-task ViLBERT on grounding based on the GQA dataset (Hudson and Manning, 2019). The grounding score shows how successful the model attends to regions within the image that are relevant to the question. We evaluate the grounding score on the *val* set, as there is no scene graph annotation available for the *testdev* set. In addition to the overall grounding scores, we also show the grounding scores for 8 attention heads of multi-task VilBERT. Asterisk denotes the best performance.

5.4 Ablations of Fusion Method

We present ablations of fusion method for both MCAN and ViLBERT in Table 4 and 5. For ease of comparison, we also include the result with unmodified fusion method (element-wise product for multi-task ViLBERT and element-wise sum for MCAN). As expected, for multi-task ViLBERT, element-wise sum does not perform well as it is trained with element-wise product strategy for both VQA 2.0 and GQA. Surprisingly, the performance of ViLBERT on the yes/no question category of VQA 2.0 only degrades by 1.49 points. This shows that questions with binary answers in VQA 2.0 are 'easier' to solve compared to other question categories. The 'other' question category in VQA 2.0 stands out as being specifically hard to solve for multi-task ViLBERT with element-wise sum achieving a very low score of only 5.94 points.

We lack sufficient evidence to suggest whether element-wise sum or element-wise product is better for cross-modal feature fusion, as the result for MCAN is contradictory for VQA 2.0 and GQA. We observe that while for VQA 2.0 using element-wise sum results in better performance, for GQA it is the opposite. The only exception occurs when we apply BERT-large as features with element-wise product for GQA, which might be due to the low dimensionality of the LSTM encoder.

5.5 Qualitative Analysis

We provide a qualitative comparison of visual grounding between MCAN and multi-task ViLBERT on the GQA *testdev* set. In order to do this, we use the learned attention from our best-performing MCAN model and multi-task ViLBERT. We did not include comparison for VQA 2.0 as it is highly unreliable due to strong priors which are exploited by the models for inference, instead of relying on visual understanding (Zhang et al., 2016; Agrawal et al., 2018).

For MCAN, we visualize the learned attentions in the image attention reduction module (Eq. (10)). We highlight the image region where the model focuses its attention the most and blur out the area where the attention score is low. As ViLBERT does not have the attention reduction module, we

Model	Overall	Y/N	Num	Other
MCAN /w BERT-base features +	70.70	86.82	52.94	60.96
MCAN /w BERT-base features ⊙	70.28	86.35	52.71	60.53
MCAN /w BERT-large features +	70.53	86.51	53.01	60.86
MCAN /w BERT-large features ⊙	70.11	86.25	52.90	60.22
MCAN /w BERT-base encoder +	71.03	87.34	53.30	61.12
MCAN /w BERT-base encoder ⊙	70.51	87.00	53.15	60.35
Multi-task ViLBERT ⊙	72.57	88.33	56.72	62.68
Multi-task ViLBERT +	43.72	86.84	45.20	5.94

Table 4: Comparison of performance between MCAN and ViLBERT with element-wise sum and element-wise product fusion on VQA 2.0. The unmodified methods are denoted with **bold**. + denotes element-wise sum while ⊙ denotes element-wise product.

Model	Accuracy	Binary	Open	Distribution
MCAN /w BERT-base features +	54.79	72.96	39.37	1.70
MCAN /w BERT-base features ⊙	55.76	73.76	40.50	1.55
MCAN /w BERT-large features +	54.70	72.18	39.87	1.74
MCAN /w BERT-large features ⊙	54.33	72.18	39.19	1.76
MCAN /w BERT-base encoder +	53.62	71.12	38.77	1.77
MCAN /w BERT-base encoder ⊙	54.14	71.80	39.16	1.58
Multi-task ViLBERT ⊙	59.75	78.49	43.85	1.31
Multi-task ViLBERT +	42.58	67.63	21.34	8.32

Table 5: Comparison of performance between MCAN and ViLBERT with element-wise sum and element-wise product fusion on GQA. The unmodified methods are denoted with **bold**. + denotes element-wise sum while ⊙ denotes element-wise product.

visualize the learned attention in a different manner. We use attention head 7 from the last layer, as it performs the best in §5.3 and visualize the sentence to image co-attention. For each concept/object that exists in the image, we take the 10 most attended regions for that concept/object and show the corresponding image patches. Figure 6 shows several examples of learned attentions for MCAN and multi-task ViLBERT.

From the examples, we can see the difference of grounding capability between MCAN and multi-task ViLBERT. Although ViLBERT performs better compared to MCAN in GQA, it does not actively focus on the concept/object relevant for answering the question most of the time (e.g., in the first question-image pair the attention maps for 'clothing' do not focus on the correct object in the image, and in the third example, the attention maps for 'small side table' do not focus on any of the tables). Occasionally, ViLBERT demonstrates the capability to learn meaningful visual grounding. This happens in the last example, where ViLBERT is able to show the relation between 'food' and 'cake' and their corresponding image regions, although it outputs the incorrect prediction. We hypothesize that for 'easy' questions, ViLBERT does not actively ground its reasoning in images as much as it should while for the harder ones, ViLBERT focuses on the region that is relevant for answering the question. This is possibly caused by ViLBERT utilizing other attention heads for 'easier' questions, and using this head for questions that require fine-grained recognition.

On the other hand, MCAN consistently grounds its reasoning in the image. For both correct and incorrect predicted examples, the attention maps focus on the most relevant image regions for answering the question. For instance, in the third example MCAN mostly attends to the left side of the room, with a particular focus on the table. This is also followed by predicting 'left' for the answer, which is consistent with the attention maps albeit incorrect.

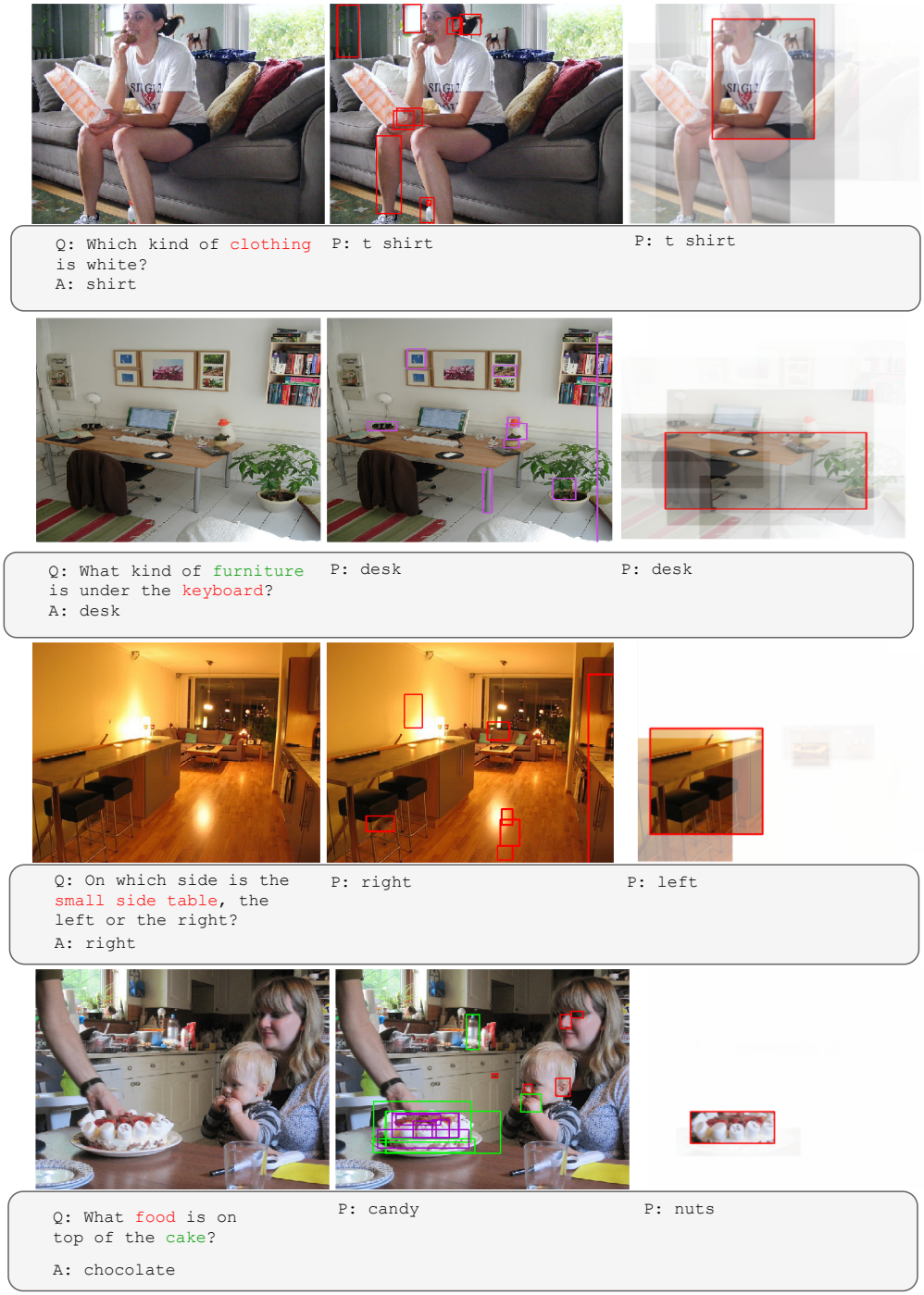


Figure 6: Examples of attention visualization for MCAN and ViLBERT on GQA *testdev* set. From left to right are the original image-question pair, prediction and the learned image attention for multi-task ViLBERT, prediction and the learned image attention for MCAN. For ViLBERT, the bounding boxes correspond to the sentence to image attention of the highlighted words in the question. We use a different bounding box color when the attention is shared between different words. As for MCAN, we visualize the attention learned by Eq. (10) and highlight the region with the highest attention score.

6 Conclusion and Future Work

In this work, we investigate the properties of joint multimodal representations derived from both a task-specific model and a multi-task model with respect to different training objective and information streams. We compare MCAN and multi-task ViLBERT on the VQA task and evaluate their performance on the VQA 2.0 and GQA datasets. The results give us an insight into the diverse improvements of the joint multimodal representations induced by the different architectures.

We show that on the one hand, ViLBERT improves the representations due to its bidirectional information streams and multi-task training, which also has a regularizing effect on the model and helps with learning underlying associations between language and visual concepts (e.g., for relations that rarely appear in VQA tasks). Although this results in high accuracy, we also demonstrate that multi-task ViLBERT does not actively ground its reasoning in images through the GQA grounding metric. On the other hand, we postulate that MCAN improves the representations by summarizing the information in various attention heads with the attention reduction module, which helps the model to refocus on the most important information and thus improves conceptual grounding. Furthermore, we observed only a modest increase in performance when applying BERT to MCAN, which indicates that the substantial improvement that BERT offers to various NLP tasks might not necessarily translate to language and vision tasks.

For future work, it would be interesting to further exploit the bidirectional characteristic of BERT embeddings by passing them through a bidirectional LSTM to obtain better question representations. In addition, it might also be worth to explore other contextual word representations derived from better language models such as RoBERTa (Liu et al., 2019). Evaluating the models on more VQA datasets such as TDIUC (Kafle and Kanan, 2017) and VQA-CP (Agrawal et al., 2018), which we briefly discussed in §2.2, could also give us a deeper insight into the properties of representations specifically in regards to image reasoning and understanding. The results from VQA-CP would be especially useful to ascertain whether the underlying associations between language and visual concepts learned by multi-task ViLBERT are truly grounded, as the answer distributions are extremely dissimilar between training and validation sets. Lastly, applying Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017) for the visualization of attention maps to our models would allow us to compare their attention maps to human attention maps via rank correlation (Das et al., 2016).

References

A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 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 (CVPR)*, June 2018.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In S. Dasgupta and D. McAllester, editors, *Proceedings of Machine Learning Research*, volume 28, pages 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/andrew13.html.

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

P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, apr 2010. doi: 10.1007/s00530-010-0182-0. URL https://doi.org/10.1007%2Fs00530-010-0182-0.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016.

T. Baltrušaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

M. Baroni. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016. doi: 10.1111/lnc3.12170. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12170.

L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660, 1999. doi: 10.1017/S0140525X99002149.

L. W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645, 2008. doi: 10.1146/annurev.psych.59.103006.093639. URL https://doi.org/10.1146/annurev.psych.59.103006.093639. PMID: 17705682.

L. Beinborn, T. Botschen, and I. Gurevych. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA, aug 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1197.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

E. Bruni, G. B. Tran, and M. Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, page 22–32, USA, 2011. Association for Computational Linguistics. ISBN 9781937284169.

S. S. Bucak, R. Jin, and A. K. Jain. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2014.

J. Chen, Z. Chen, Z. Chi, and H. Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 508–513, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328852. doi: 10.1145/2663204.2666277. URL https://doi.org/10.1145/2663204.2666277.

S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pages 49–56, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337434. doi: 10.1145/2808196.2811638. URL https://doi.org/10.1145/2808196.2811638.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas, nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1092. URL https://www.aclweb.org/anthology/D16-1092.

H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3), Feb. 2015. ISSN 0360-0300. doi: 10.1145/2682899. URL https://doi.org/10.1145/2682899.

G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, nov 2013. doi: 10.1109/tmm.2013.2267205. URL https://doi.org/10.1109%2Ftmm.2013.2267205.

R. Fincher-Kiefer. Perceptual components of situation models. *Memory & Cognition*, 29:336–343, 2001. URL https://doi.org/10.3758/BF03194928.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

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 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1044. URL https://www.aclweb.org/anthology/D16-1044.

H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2296–2304. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5641-are-you-talking-to-a-machine-dataset-and-methods-for-multilingual-image-question.pdf.

P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li. 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 (CVPR)*, pages 6639–6648, June 2019.

P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 221–228, 2009.

D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1422953112. URL https://www.pnas.org/content/112/12/3618.

M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker. Multiple classifier systems for the classification of audio-visual emotional states. In S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, editors, *Affective Computing and Intelligent Interaction*, pages 359–368, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24571-8.

M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12(null): 2211–2268, July 2011. ISSN 1532-4435.

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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017.

H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443 Vol. 4, 2005.

M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit. Dynamic modality weighting for multi-stream hmms in audio-visual speech recognition. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 237–240, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581989. doi: 10.1145/1452392.1452442. URL https://doi.org/10.1145/1452392.1452442.

S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990. URL http://cogprints.org/3106/.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

F. Hill, R. Reichart, and A. Korhonen. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014. doi: 10.1162/tacl_a_00183. URL https://www.aclweb.org/anthology/Q14-1023.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *International Conference on Learning Representations (ICLR)*, 2018.

D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

N. Jaques, S. Taylor, A. Sano, and R. Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *NIPS Workshop on Multimodal Machine Learning*, Montreal, Quebec, 2015. URL https://www.media.mit.edu/publications/multi-task-multi-kernel-learning-for-estimating-individual-wellbeing/.

Q. Jin and J. Liang. Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 239–242, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343596. doi: 10.1145/2911996.2912043. URL https://doi.org/10.1145/2911996.2912043.

M. Joshi, O. Levy, L. Zettlemoyer, and D. Weld. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1588. URL https://www.aclweb.org/anthology/D19-1588.

K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1983–1991, 2017.

S. Kahou, X. Bouthillier, P. Lamblin, Çaglar Gülçehre, V. Michalski, K. R. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. C. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10:99–111, 2015.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://www.aclweb.org/anthology/D14-1086.

J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1564–1574. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7429-bilinear-attention-networks.pdf.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. URL https://doi.org/10.1007/s11263-016-0981-7.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia Tools and Applications*, 71:333–347, 2014. URL https://doi.org/10.1007/s11042-013-1391-2.

C. W. Leong and R. Mihalcea. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1403–1407, Chiang Mai, Thailand, nov 2011. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I11-1162.

L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.469. URL https://www.aclweb.org/anthology/2020.acl-main.469.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 289–297. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering.pdf.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input.pdf.

M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015.

M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision*, 125:110–135, 2017. URL https://doi.org/10.1007/s11263-017-1038-2.

S. Manmadhan and B. Kovoor. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, 04 2020. doi: 10.1007/s10462-020-09832-7.

J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013a. URL http://arxiv.org/abs/1301.3781. cite arxiv:1301.3781.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013b. Curran Associates Inc.

G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 1991.

E. Morvant, A. Habrard, and S. Ayache. Majority vote of diverse classifiers for late fusion. In P. Fränti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 153–162, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-662-44415-3.

Y. Mroueh, E. Marcheret, and V. Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134, 2015.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *ICML*, 2011.

D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2): 92–105, 2011.

B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, pages 284–288, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345569. doi: 10.1145/2993148.2993176. URL https://doi.org/10.1145/2993148.2993176.

Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4594–4602. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.497. URL https://doi.org/10.1109/CVPR.2016.497.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 2641–2649, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

S. Poria, E. Cambria, and A. Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1303. URL https://www.aclweb.org/anthology/D15-1303.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

T. Qin, T. yan Liu, X. dong Zhang, D. sheng Wang, and H. Li. Global ranking using continuous conditional random fields. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1281–1288. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3402-global-ranking-using-continuous-conditional-random-fields.pdf.

A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1848–1852, Oct. 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1124. URL https://doi.org/10.1109/TPAMI.2007.1124.

S. S. Rajagopalan, L.-P. Morency, T. Baltrušaitis, and R. Goecke. Extending long short-term memory for multi-view structured learning. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 338–353, Cham, 2016. Springer International Publishing.

G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 396–406, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642245701.

M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2953–2961. Curran Associates, Inc., 2015a. URL http://papers.nips.cc/paper/5640-exploring-models-and-data-for-image-question-answering.pdf.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015b. URL http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In D. van Dyk and M. Welling, editors, *Proceedings of Machine Learning Research*, volume 5, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/salakhutdinov09a.html.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://www.aclweb.org/anthology/P18-1238.

E. Shutova, D. Kiela, and J. Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San

Diego, California, jun 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1020. URL https://www.aclweb.org/anthology/N16-1020.

K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 517–524, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321297. doi: 10.1145/2522848.2531741. URL https://doi.org/10.1145/2522848.2531741.

C. Silberer and M. Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland, jun 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1068. URL https://www.aclweb.org/anthology/P14-1068.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *international conference on learning representations*, 2015.

A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 399–402, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930442. doi: 10.1145/1101149.1101236. URL https://doi.org/10.1145/1101149.1101236.

Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: Learning correlation and interaction across modalities. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 27–30, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314671. doi: 10.1145/2388676.2388684. URL https://doi.org/10.1145/2388676.2388684.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://www.aclweb.org/anthology/P19-1644.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, Jan. 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*,

pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *Proc. Int. Conf. Learn. Representations*, 2016.

S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving LSTM-based video description with linguistic knowledge mined from text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1204. URL https://www.aclweb.org/anthology/D16-1204.

L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5005–5013, 2016.

P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427, 2018.

J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 2764–2770. AAAI Press, 2011. ISBN 9781577355151.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019.

J. Wu and R. Mooney. Self-critical reasoning for robust visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8604–8614. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9066-self-critical-reasoning-for-robust-visual-question-answering.pdf.

Z. Wu, L. Cai, and H. Meng. Multi-level fusion of audio and visual features for speaker identification. In D. Zhang and A. K. Jain, editors, *Advances in Biometrics*, pages 493–499, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31621-3.

Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *MM '14*, 2014.

N. Xie, F. Lai, D. Doran, and A. Kadav. Visual entailment task for visually-grounded language learning, 2018.

N. Xie, F. Lai, D. Doran, and A. Kadav. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706, 2019. URL http://arxiv.org/abs/1901.06706.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 451–466, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

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

Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C. F. Wang. A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection. *IEEE Transactions on Multimedia - TMM*, 14:563–574, 06 2012. doi: 10.1109/TMM.2012.2188783.

T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018.

Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2019.

P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022, 2016.

Y. Zhang, J. C. Niebles, and A. Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357, 2019.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.11. URL https://doi.org/10.1109/ICCV.2015.11.

Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.