Fusion of Detected Objects in Text for Visual Question Answering

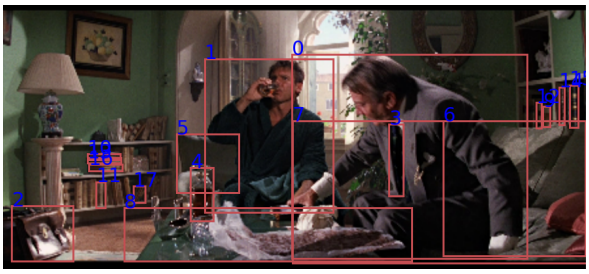
Chris Alberti **Jeffrey Ling*** **Michael Collins** **David Reitter**
Google Research
{chrisalberti, jeffreyling, mjcollins, reitter}@google.com

Abstract

To advance models of multimodal context, we introduce a simple yet powerful neural architecture for data that combines vision and natural language. The "Bounding Boxes in Text Transformer" (B2T2) also leverages referential information binding words to portions of the image in a single unified architecture. B2T2 is highly effective on the Visual Commonsense Reasoning benchmark[1], achieving a new state-of-the-art with a 25% relative reduction in error rate compared to published baselines and obtaining the best performance to date on the public leaderboard (as of May 22, 2019). A detailed ablation analysis shows that the early integration of the visual features into the text analysis is key to the effectiveness of the new architecture. A reference implementation of our models is provided[2].

1 Introduction

It has long been understood that the meaning of a word is systematically and predictably linked to the context in which it occurs (e.g., Firth 1957; Harris 1954; Deerwester et al. 1990; Mikolov et al. 2013). Different notions of context have resulted in different levels of success with downstream NLP tasks. Recent neural architectures including Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2018) have dramatically increased our ability to include a broad window of potential lexical hints. However, the same capacity allows for multimodal context, which may help model the meaning of words in general, and also sharpen its understanding of instances of words in context (e.g., Bruni et al. 2014).



Q: What was [1] doing before he sat in his living room?
A_1: He was reading [10].
A_2: He was taking a shower. ✓
A_3: [0] was sleeping until the noise [1] was making woke him up.
A_4: He was sleeping in his bedroom.
R_1: His clothes are disheveled and his face is glistening like he's sweaty.
R_2: [0] does not look wet yet, but [0] looks like his hair is wet, and bathrobes are what you wear before or after a shower.
R_3: He is still wearing his bathrobe. ✓
R_4: His hair appears wet and there is clothing hanging in front of him on a line as if to dry.

Figure 1: An example from the VCR dataset. The tasks consists in picking an answer A_{1-4}, and then picking a rationale R_{1-4}. The data contains explicit pointers in the text to bounding boxes in the image.

In this paper, we consider visual context in addition to language and show that the right integration of visual and linguistic information can yield improvements in visual question answering. The challenge we consider is to answer natural-questions related to a given image. The more general question we address in the context of this problem is how to encode visual and verbal information in a neural architecture. How to best do that is still unclear. How are text entities bound to objects seen in images? Are text and image best integrated late, allowing for independent analysis (*late fusion*), or should the processing of one be conditioned on the analysis of the other (*early fusion*)? How is cross-modal co-reference best encoded at all? Does it make sense to ground words in the visual world before encoding sentence semantics?

*Work done as part of the Google AI residency.
[1] https://visualcommonsense.com
[2] https://github.com/google-research/language/tree/master/language/question_answering/b2t2

In this work we gather evidence to answer these questions by designing the Bounding Boxes in Text Transformer, B2T2 for short, a neural architecture for multimodal encoding of natural language and images, and we evaluate B2T2 on the Visual Commonsense Reasoning benchmark (VCR, Zellers et al. 2019).

Figure 1 shows an illustrative example from the VCR benchmark. VCR is well suited to test rich multimodal representations because it requires the analysis of images depicting people engaged in complex activities; it presents questions, answers and rationales created by human annotators rather than automatic generation; it has a clean multiple-choice interface for evaluation; and yet it is still challenging thanks to a careful selection of answer choices through adversarial matching. VCR has much longer questions and answers compared to other popular Visual Question Answering (VQA) datasets, such as VQA v1 (Antol et al., 2015), VQA v2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019), requiring more modeling capacity for language understanding.

In our experiments, we found that early fusion of co-references between textual tokens and visual features of objects was the most critical factor in obtaining improvements on VCR. We found that the more visual object features we included in the model's input, the better the model performed, even if they were not explicitly co-referent to the text, and that positional features of objects in the image were also helpful. We finally discovered that our models for VCR could be trained much more reliably when they were initialized from pre-training on Conceptual Captions (Sharma et al., 2018), a public dataset of about 3M images with captions. From the combination of these modeling improvements, we obtained a new model for visual question answering that achieves state-of-the-art on VCR, reducing error rates by more than 25% relative to the best published and documented model (Zellers et al., 2019).

2 Problem Formulation

In this work, we assume data comprised of 4-tuples (I, B, T, l) where

1. I is an image,

2. $B = [b_1, \ldots, b_m]$ is a list of bounding boxes referring to regions of I, where each b_i is identified by the lower left corner, height and width,

3. $T = [t_1, \ldots, t_n]$ is a passage of tokenized text, with the peculiarity that some of the tokens are not natural language, but explicit references to elements of B, and

4. l is a binary label in $\{0, 1\}$.

Symbol	Type	Description
m	$\mathbb{N}$	number of extracted bounding boxes
n	$\mathbb{N}$	number of tokens input to BERT
k	$\mathbb{N}$	number of positional embeddings for image coordinates, usually 56
d	$\mathbb{N}$	visual features dimension, usually 2048
h	$\mathbb{N}$	hidden dimension of BERT, usually 1024
l	$\{0, 1\}$	a binary label
I	$\mathbb{R}^{\cdot \times \cdot \times 3}$	an image
B	$\mathbb{R}^{m \times 4}$	rectangular bounding boxes on I, as coordinates of opposite corners
R	$\{0, 1\}^{m \times n}$	matrix encoding which bounding boxes in B correspond to which tokens in T
T	$\mathbb{N}^{n \times 2}$	input tokens, each expressed as word piece id and token type
Φ	$\mathbb{R}^{\cdot \times \cdot \times 3} \to \mathbb{R}^d$	a function to extract visual feature vectors from an image
π	$\mathbb{R}^4 \to \mathbb{R}^d$	a function to embed the position and shape of a bounding box
Ψ	$\mathbb{R}^{n \times h} \to \mathbb{R}^h$	a function to compute a passage embedding from per-token embeddings
E	$\mathbb{N}^{n \times 2} \to \mathbb{R}^{n \times h}$	non-contextualized token embeddings, encoding word piece ids, token types and positions

Table 1: Glossary of mathematical symbols used in this work.

While it might seem surprising to mix natural text with explicit references to bounding boxes, this is actually a quite natural way for people to discuss objects in images and the VCR dataset is annotated in exactly this way.

We assume an image representation function Φ that converts an image, perhaps after resizing and padding, to a fixed size vector representation of dimension d.

We similarly assume a pretrained textual representation capable of converting any tokenized passage of text, perhaps after truncating or padding, into a vector representation of dimension h. We assume a context independent token representation E in the shape of a vector of dimension h for each token and a passage level representation Ψ which operates on $E(T)$ and returns a passage level vector representation of dimension h.

We refer the reader to Table 1 for an overview of the notation used in this work. Full details on how the VCR dataset is encoded into this formalism are given in Section 4.

3 Models and Methods

We evaluate two main architectures: "Dual Encoder", a late fusion architecture where image and text are encoded separately and answer scores are computed as an inner product, and the full B2T2 model, an early fusion architecture where visual features are embedded on the same level as input word tokens. Section 5.2 will summarize experiments with model variants to answer the research questions laid out in the introduction and to analyze what works, and why.

3.1 Dual Encoder

Dual Encoders, discussed for example by Wu et al. (2018) and Gillick et al. (2018), are models that embed objects of potentially different types into a common representation space where a similarity function can be expressed e.g. as a dot product or a cosine similarity. A notable example of a dual encoder for image classification is WSABIE, proposed by Weston et al. (2011).

Our Dual Encoder architecture is shown in Figure 2. We model the class distribution as

$$p(l = 1|I, T) = \frac{1}{1 + e^{-\Psi(E(T))^\top D\Phi(I)}}$$

where D is a learned matrix of size $d \times h$. In this model, co-reference information is completely ignored, and the model must rely on fixed dimensional vectors for the late fusion of textual and visual contexts. However, we found this to be surprisingly competitive on VCR compared to published baselines, perhaps due to our choice of powerful pretrained models.

3.2 B2T2

Our B2T2 architecture is shown in Figure 3. We model the class distribution as

$$p(l|I, B, R, T) = \frac{e^{\Psi(E'(I,B,R,T))\cdot a_l + b_l}}{\sum_{l'} e^{\Psi(E'(I,B,R,T))\cdot a_{l'} + b_{l'}}}$$

where $a_l \in \mathbb{R}^h$ and $b_l \in \mathbb{R}$ for $l \in \{0, 1\}$ are learned parameters. $E'(I, B, R, T)$ is a non-contextualized representation for each token and of its position in text, but also of the content and position of the bounding boxes. The key difference from "Dual Encoder" is that text, image and bounding boxes are combined at the level of the non-contextualized token representations rather than right before the classification decision.

The computation of $E'(I, B, R, T)$ is depicted in Figure 4. More formally, for a given example, let matrix $R \in \{0, 1\}^{m\times n}$ encode the references between the bounding boxes in B and the tokens in T, so that R_{ij} is 1 if and only if bounding box i is referenced by token j. Then

$$E'(I, B, R, T) = E(T) + \sum_{i=1}^{m} R_i \left[M(\Phi(\text{crop}(I, b_i)) + \pi(b_i))\right]^\top$$

where M is a learned $h \times d$ matrix, $\Phi(\text{crop}(I, b_i))$ denotes cropping image I to bounding box b_i and then extracting a visual feature vector of size d, and $\pi(b_i)$ denotes the embedding of b_i's shape and position information in a vector of size d.

To embed the position and size of a bounding box b, we introduce two new learnable embedding matrices X and Y of dimension $k \times \frac{d}{4}$. Let the coordinates of the opposite corners of b be (x_1, y_1) and (x_2, y_2), after normalizing so that a bounding box covering the entire image would have $x_1 = y_1 = 0$ and $x_2 = y_2 = k$. Position embeddings are thus defined to be

$$\pi(b) = \text{concat}(X_{\lfloor x_1 \rfloor}, Y_{\lfloor y_1 \rfloor}, X_{\lfloor x_2 \rfloor}, Y_{\lfloor y_2 \rfloor})$$

3.3 Loss

All of our models are trained with binary cross entropy loss using label l. Denoting $p := p(l = 1|I, B, R, T)$, we have for each example

$$\mathcal{L}_{\text{BCE}} = l \log p + (1 - l) \log(1 - p)$$

3.4 Pretraining on Conceptual Captions

Before training on VCR, we pretrain B2T2 on image and caption pairs using a Mask-LM pretraining technique like the one used in BERT (Devlin et al., 2018). The setup used during pretraining is shown in Figure 5, where the model uses the image as additional context when filling in the mask.

We use two tasks for pretraining: (1) impostor identification and (2) masked language model prediction. For the impostor task, we sample a random negative caption for each image and ask the model to predict whether the caption is correctly associated. For mask-LM, we randomly replace tokens in the caption with the `[MASK]` token, and the model must predict the original token (see Devlin et al. (2018) for more details).

Formally, the pretraining data consist of images I and captions T. We do not consider bounding

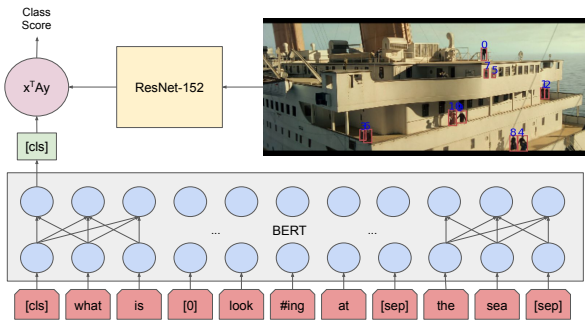


Figure 2: Dual Encoder architecture with late fusion. The model extracts a single visual feature vector from the entire image. Bounding boxes are ignored.

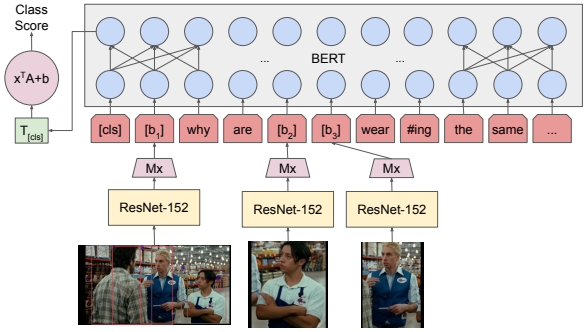


Figure 3: B2T2 architecture with early fusion. Bounding boxes are inserted where they are mentioned in the text and at the end of the input, as described in Sec. 4.

boxes during pretraining, so $B = \emptyset$. The binary label l indicates whether the caption is an impostor or not. The loss for impostor identification is binary cross entropy $\mathcal{L}_{\text{BCE}}$ with label l as in 3.3. We denote the loss for mask-LM as $\mathcal{L}_{\text{MLM}}$, which is the summed cross entropy of the predicted token distributions against the true tokens.

To ensure that our model correctly grounds the language to the image with the mask LM loss, we only use it for positive captions, zeroing it out for negative captions. Our final objective is the sum of the losses:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + I[l = 1] \cdot \mathcal{L}_{\text{MLM}}$$

where $I[l = 1]$ is an indicator for the label l being positive for the image and caption pair.

We pretrain on Conceptual Captions (Sharma et al., 2018), a dataset with over 3M images paired with captions.[3] We found empirically that pretraining improves our model slightly on VCR, but more importantly, allows our model to train stably. Without pretraining, results on VCR exhibit much higher variance. We refer the reader to Section 5.2 for an ablation analysis on the effect of pretraining.

3.5 Implementation Details

We use ResNet-152[4] (He et al., 2016) pretrained on ImageNet for Φ, which yields a vector representation of size $d = 2048$. BERT-Large (Devlin et al., 2018) provides both E and Ψ. The latter is a pretrained Transformer with 24 layers, 16 attention heads, and hidden size 1024. For BERT,

[3] We also tried pretraining on MS-COCO images and captions (Lin et al., 2014), but found this to be ineffective. This could be because MS-COCO is smaller (with around 80k images, 400k captions).

[4] Publicly available at `tfhub.dev`

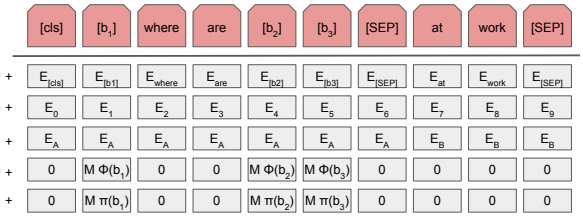


Figure 4: How input embeddings are computed in our B2T2 architecture.

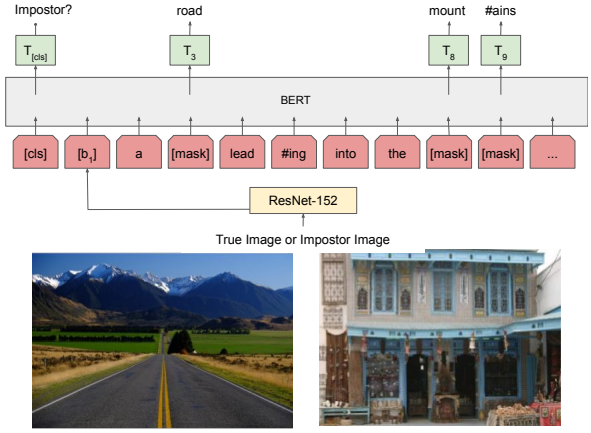


Figure 5: Mask-LM pretraining for B2T2.

E corresponds to its token embeddings, Ψ to the `[CLS]` token representation in the final layer, and so $\Psi(E(T))$ corresponds to the BERT passage representation of size $h = 1024$.

We found empirically that it was slightly better to keep Φ fixed rather than fine-tuning it, but that it was of critical importance to fine-tune Ψ and E for the new task.

In all of our finetuning experiments we use the Adam optimizer (Kingma and Ba, 2014) and trained our models with a grid of hyperparameters: a learning rate of $2 \cdot 10^{-5}$ and $3 \cdot 10^{-5}$, for 3,

4, and 5 epochs with a linear learning rate decay, and two random seed for initialization.

To maximize performance on VCR, we also evaluate an ensemble of B2T2 models. Our ensemble is comprised of 5 identical B2T2 models, trained for 3 epochs with an initial learning rate of $2 \cdot 10^{-5}$, but initialized with 5 different random seeds. The resulting class logits are then summed to obtain the ensemble scores.

4 Data

Visual Commonsense Reasoning (VCR, visualcommonsense.com, Zellers et al. 2019) is a corpus that contains a sample of stills from movies. Questions and answers revolve around conclusions or assumptions that require knowledge external to the images. The associated task is to not only select a correct answer but also provide reasoning in line with common sense. Matching our problem formulation given before, a VCR sample is defined as a tuple (I, O, Q, A, R). Here, I is the image, and O is a sequence of objects identified in the image. A question $Q = [q_0, \ldots, q_k]$ is given, where tokens are either textual words or deictic references to objects in O. Each question contains a set of four answers $A = \{A_1, A_2, A_3, A_4\}$, with exactly one correct answer A^*. Each response follows the schema of the queries. Finally, there is a set of four rationales $R = \{R_1, R_2, R_3, R_4\}$, with exactly one rationale R^* identified as correct in supporting A^*.

Each of the objects in $O = [(b_1, l_1), \ldots, (b_{|O|}, l_{|O|})]$ is identified in the image I by bounding boxes b_i. The objects are also labeled with their classes with a text token l_i.

The $Q \rightarrow A$ task is to choose A^* given (I, O, Q, A). The $QA \rightarrow R$ task is to choose R^* given (I, O, Q, A^*, R). Finally, the $Q \rightarrow AR$ task is a pipeline of the two, where a model must first correctly choose A^* from A, then correctly choose R^* given A^*.

We adapt VCR to our problem formulation by converting each VCR example to four instances for the $Q \rightarrow A$ task, one per answer in A, and four instances for the $QA \rightarrow R$ task, one per rationale in R. We construct the text for the instances in the $Q \rightarrow A$ task as

$$[\text{[CLS]}, [b_0], q_0, \ldots, \text{[SEP]}, a_0, \ldots, \text{[SEP]}, l_1, [b_1], \ldots, l_p, [b_p]]$$

and in the $QA \rightarrow R$ task as

$$[\text{[CLS]}, [b_0], q_0, \ldots, \text{[SEP]}, a_0^*, \ldots, r_0, \ldots, \text{[SEP]}, l_1, [b_1], \ldots, l_p, [b_p]].$$

where [CLS], [SEP] are special tokens for BERT.

Here, $[b_0]$ is a bounding box referring to the entire input image. $q_0, \ldots$ are all question tokens, $a_0, \ldots$ answer tokens, $a_0^*, \ldots$ answer tokens for the correct answer, and $r_0, \ldots$ rationale tokens. We append the first p bounding boxes in O with class labels to the end of the sequence (in our experiments, we use $p = 8$), and for objects referenced in Q, A, R, we prepend the class label token (i.e. $[b_i]$ becomes $l_i, [b_i]$). We assign the binary label l to every instance to represent whether the answer or rationale choice is the correct one.

5 Experimental Results

5.1 VCR Task Performance

Our final results on the VCR task are shown in Table 2. Our Dual Encoder model worked surprisingly well compared to Zellers et al. (2019), surpassing the baseline without making use of bounding boxes. We also evaluate a Text-Only baseline, which is similar to the Dual Encoder model but ignores the image. The ensemble of B2T2 models, pretrained on Conceptual Captions, obtained absolute accuracy improvements of 8.9%, 9.8% and 13.1% compared to the published R2C baseline for the $Q \rightarrow A$, $QA \rightarrow R$, and $Q \rightarrow AR$ tasks respectively. At the time of this writing (May 22, 2019), both our single B2T2 and ensemble B2T2 models outperform all other systems in the VCR leaderboard.

5.2 Ablations

To better understand the reason for our improvements, we performed a number of ablation studies on our results, summarized in Table 3. We consider ablations in order of decreasing impact on the VCR dev set $Q \rightarrow A$ accuracy.

Use of Bounding Boxes. The bounding boxes considered by our model turns out to be the most important factor in improving the accuracy of our model. Without any bounding boxes we obtain 67.5% accuracy, just above the accuracy of the dual encoder. With 4 instead of 8 appended bounding boxes we obtain 71% accuracy. With 8 bounding boxes, but no textual labels from the bounding boxes in the text we obtain 70.9% accuracy,

Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text-Only BERT (Zellers et al.)	53.8	53.9	64.1	64.5	34.8	35.0
R2C (Zellers et al.)	63.8	65.1	67.2	67.3	43.1	44.0
HCL HGP (unpub.)	-	70.1	-	70.8	-	49.8
TNet (unpub.)	-	70.9	-	70.6	-	50.4
B-VCR (unpub.)	-	70.5	-	71.5	-	50.8
TNet 5-Ensemble (unpub.)	-	72.7	-	72.6	-	53.0
Text-Only BERT (ours)	59.5	-	65.6	-	39.3	-
Dual Encoder (ours)	66.8	-	67.7	-	45.3	-
B2T2 (ours)	71.9	72.6	76.0	75.7	54.9	55.0
B2T2 5-Ensemble (ours)	**73.2**	**74.0**	**77.1**	**77.1**	**56.6**	**57.1**
Human		91.0		93.0		85.0

Table 2: Experimental results on VCR, incorporating those reported by Zellers et al. (2019). The proposed B2T2 model and the B2T2 ensemble outperform published and unpublished/undocumented results found on the VCR leaderboard at `visualcommonsense.com/leaderboard` as of May 22, 2019.

	$Q \rightarrow A$
Dual Encoder	66.8
No bboxes	67.5
Late fusion	68.6
BERT-Base	69.0
ResNet-50	70.4
No bbox class labels	70.9
Fewer appended bboxes ($p = 4$)	71.0
No bbox position embeddings	71.6
Full B2T2	71.9

Table 3: Ablations for B2T2 on VCR dev. The Dual Encoder and the full B2T2 models are the main models discussed in this work. All other models represent ablations from the full B2T2 model.

showing that our model can make use of labels for detected objects. Example 1 in Table 4 shows an example that our models can only get right if bounding box 5 is available.

Late Fusion vs. Early Fusion. The second most important architectural choice in our model is to combine visual information at the level of context independent token embeddings, rather than at the highest levels of the neural representation. If in the the full B2T2 model we add visual embeddings in the last layer of BERT rather than in the first, we lose 3.3% accuracy.

Effect of Textual Model Size. The original VCR work by Zellers et al. (2019) made use of BERT-base, while we use BERT-large to initialize our models. To test how much of our improvements are simply due to our model being larger, we retrained B2T2 models using BERT-base and found that we lose 2.9% accuracy.

Effect of Visual Model Size. How important is the choice of the visual model in the performance

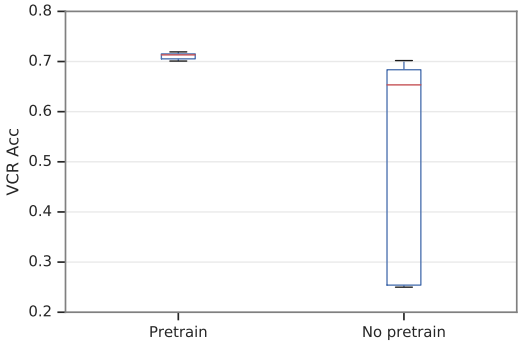


Figure 6: Boxplot of dev $Q \rightarrow A$ accuracy on VCR with and without pretraining. Pretraining on Conceptual Captions lowers variance when fine-tuning on VCR, from a grid search on multiple random seeds, learning rates, and VCR training epochs.

of B2T2? As further discussed in the error analysis section of this work, we suspect that B2T2 could be significantly improved by extending the visual features to represent more than just objects, but also activities, expressions and more. However it appears that even the size of the object detection model is important. If we swap out ResNet-152 for ResNet-50, accuracy decreases by 1.5%.

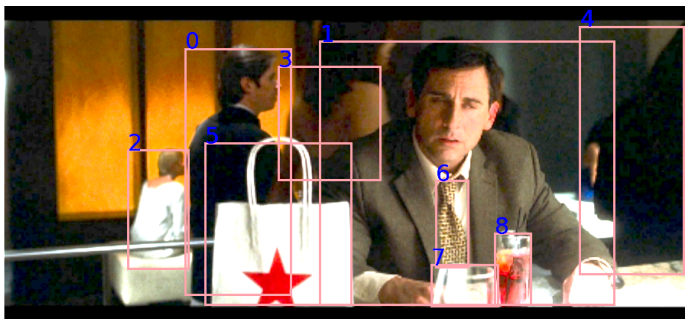
Pretraining. We found that performance improvements from pretraining are quite small, around 0.4% accuracy, but initializing from a pretrained model heavily reduces variance of results. We show this effect in Figure 6 over the grid of learning rates, random seeds, and training epochs described in Section 3.5.

Position of Bounding Boxes We additionally investigated the effect of removing position information from the model. The benefit of having bounding box positional embeddings is the smallest of the ones we considered. A model trained without positional embeddings only loses 0.3% accuracy compared to the full model.

5.3 Error Analysis

We picked some examples, shown in Table 4, to illustrate the kinds of correct and incorrect choices that B2T2 is making, compared to our dual encoder and to a text only model.

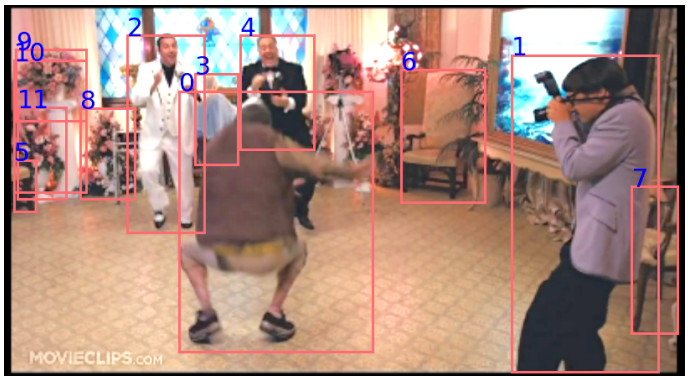
In Example 1 we show an example of how our model picks the right answer only when it is able to make use of all provided bounding boxes. Bounding box 5 in particular contains the clue that allows the observer to know that the man in the picture might have just gone shopping.



Example 1

Q: What did [1] do before coming to this location?

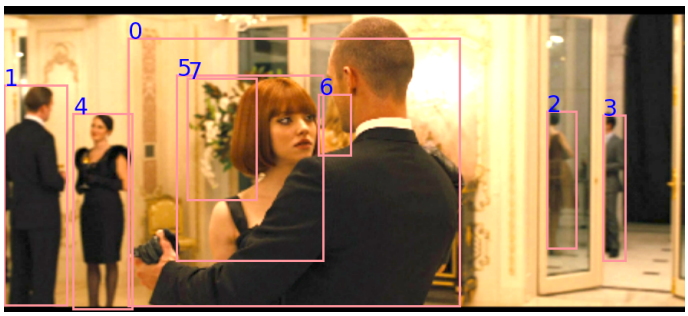
A_1: He took horse riding lessons. (text-only)
A_2: He was just shopping. (B2T2)
A_3: He found a skeleton.
A_4: He came to buy medicine. (dual encoder)



Example 2

Q: How are [2, 4] related?

A_1: [2, 4] are partners on the same mission.
A_2: [2, 4] are a recently married gay couple. (B2T2)
A_3: They are likely acquaintances.
A_4: They are siblings. (text-only, dual encoder)



Example 3

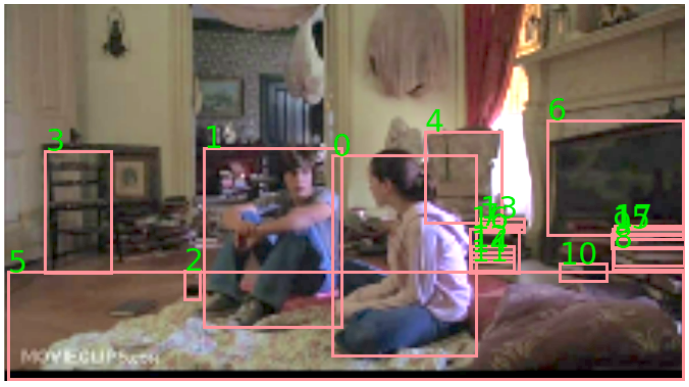
Q: What are [0] and the woman doing?

A_1: Their husbands are doing something dumb.
A_2: They are observing the results of an experiment. (text-only, dual encoder)
A_3: They are dancing. (B2T2)
A_4: They are acting as nurses for the rescued people.

Example 4

Q: How is [2] feeling?

A_1: [2] is feeling shocked. (B2T2, dual encoder)
A_2: [0] is feeling anxious.
A_3: [2] is not feeling well.
A_4: [2] is feeling joy and amusement. (text-only)



Example 5

Q: Why is [1] on the floor talking to [0]?

A_1: The man on the floor was assaulting [1].
A_2: He is asking her to help him stand up. (B2T2, dual encoder)
A_3: [1] just dropped all his books on the floor.
A_4: [1] looks like he is telling [0] a secret. (text-only)

Table 4: Examples of the $Q \rightarrow A$ task from the VCR dev set. The correct answer for every example is marked in bold. The answers picked by the text-only model, by the dual encoder and by B2T2 are indicated in parenthesis.

In Examples 2 and 3, no specific bounding box appears to contain critical clues for answering the question, but B2T2 outperforms models without access to the image or without access to bounding boxes. It is possible that B2T2 might be gaining deeper understanding of a scene by combining information from important regions of the image.

In Examples 4 and 5, we see failure cases of both the dual encoder and B2T2 compared to the text only-model. Both these examples appear to point to a limitation in the amount of information that the we are able to extract from the image. Indeed our vision model is trained on ImageNet, and so it might be very good at recognizing objects, but might be unable to recognize human expressions and activities. Our models could have correctly answered the question in Example 4 if they were able to recognize smiles. Similarly our models could have ruled out the incorrect answer they picked for the question in Example 5 if they were able to see that both people in the picture are sitting down and are not moving.

6 Related Work

Modeling visual contexts can aid in learning useful sentence representations (Kiela et al., 2017) and even in training language models (Ororbia et al., 2019). This paper takes these more general ideas to a downstream task that requires modeling of visual input. Similar to B2T2, VideoBERT (Sun et al., 2019) jointly processes video frames and text tokens with a Transformer architecture. However, VideoBERT cannot answer questions, nor does the model consider bounding boxes.

Our B2T2 model is similar to the Bottom-Up Top-Down attention model (Anderson et al., 2018) in how bounding boxes generated at preprocessing time are attended to by the VQA model. "Bottom-Up" refers to the idea of attending from the text to the bounding boxes of objects detected in the image, while "Top-Down" refers to the idea of attending to regions constructed as a regular grid over the image. The Bottom-Up Top-Down model however reduces the text to a fixed length vector representation before attending to image regions, while B2T2 instead treats image regions as special visual tokens mixed in the text. In this sense, Bottom-Up Top-Down model is a late fusion model, while B2T2 is early fusion.

The Neuro-Symbolic Concept Learner (Mao et al., 2019) also uses bounding boxes to learn visually grounded concepts through language. The Neuro-Symbolic Concept Learner however relies on a semantic parser to intepret language, while B2T2 uses a Transformer to construct a joint representation of textual tokens and visual tokens.

Another recently proposed model for VQA is MAC (Hudson and Manning, 2018). As presented, MAC does not make use of bounding boxes, which makes it a Top-Down model in the nomenclature of Anderson et al. (2018). MAC also reduces the textual information to a vector of fixed length. However MAC makes use of a new neural architecture designed to perform an explicit multi-step reasoning process and is reported to perform better than Anderson et al. (2018) on the GQA dataset (Hudson and Manning, 2019).

After the submission of this paper, several new works were published with excellent results on VCR, in some cases exceeding the performance of our system. In particular we mention ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2019), Unicoder-VL (Li et al., 2019a), and VisualBERT (Li et al., 2019b).

VCR is only one of several recent datasets pertaining to the visual question answering task. VQA (Antol et al., 2015; Zhang et al., 2016; Goyal et al., 2017) contains photos and abstract scenes with questions and several ground-truth answers for each, but the questions are less complex than VCR's. CLEVR (Johnson et al., 2017) is a visual QA task with compositional language, but the scenes and language are synthetic. GQA (Hudson and Manning, 2019) uses real scenes from Visual Genome, but the language is artificially generated. Because VCR has more complex natural language than other datasets, we consider it the best evaluation of a model like B2T2, which has a powerful language understanding component.

7 Conclusion

In this work we contrast different ways of combining text and images when powerful text and vision models are available. We picked BERT-Large (Devlin et al., 2018) as our text model, ResNet-152 (He et al., 2016) as our vision model, and the VCR dataset (Zellers et al., 2019) as our main benchmark.

The early-fusion B2T2 model, which encodes sentences along with links to bounding boxes around identified objects in the images, produces the best available results in the visual question an-

swering tasks. A control model, implementing late fusion (but the same otherwise), performs substantively worse. Thus, grounding words in the visual context should be done early rather than late.

We also demonstrate competitive results with a Dual Encoder model, matching state-of-the-art on the VCR dataset even when textual references to image bounding boxes are ignored. We then showed that our Dual Encoder model can be substantially improved by deeply incorporating in the textual embeddings visual features extracted from the entire image and from bounding boxes. We finally show that pretraining our deep model on Conceptual Captions with a Mask-LM loss yields a small additional improvement as well as much more stable fine-tuning results.

References

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. In *ICLR*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2017. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Alexander G. Ororbia, Ankur Mali, Matthew A. Kelly, and David Reitter. 2019. Like a baby: Visually situated neural language acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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*, pages 2556–2565.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.