Cross-Modality Relevance for Reasoning on Language and Vision

Chen Zheng, Quan Guo, Parisa Kordjamshidi
Department of Computer Science and Engineering, Michigan State University
East Lansing, MI 48824
{zhengc12,guoquan,kordjams}@msu.edu

Abstract

This work deals with the challenge of learning and reasoning over language and vision data for the related downstream tasks such as *visual question answering (VQA)* and *natural language for visual reasoning (NLVR)*. We design a novel cross-modality relevance module that is used in an end-to-end framework to learn the relevance representation between components of various input modalities under the supervision of a target task, which is more generalizable to unobserved data compared to merely reshaping the original representation space. In addition to modeling the relevance between the textual entities and visual entities, we model the higher-order relevance between entity relations in the text and object relations in the image. Our proposed approach shows competitive performance on two different language and vision tasks using public benchmarks and improves the state-of-the-art published results. The learned alignments of input spaces and their relevance representations by NLVR task boost the training efficiency of VQA task.

1 Introduction

Real-world problems often involve data from multiple modalities and resources. Solving a problem at hand usually requires the ability to reason about the components across all the involved modalities. Examples of such tasks are visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017) and natural language visual reasoning (NLVR) (Suhr et al., 2017, 2018). One key to intelligence here is to identify the relations between the modalities, combine and reason over them for decision making. Deep learning is a prominent technique to learn representations of the data for decision making for various target tasks. It has achieved supreme performance based on large scale corpora (Devlin et al., 2019). However, it is a challenge to learn joint representations for cross-modality data because deep learning is data-hungry. There are many recent efforts to build such multi-modality datasets (Lin et al., 2014; Krishna et al., 2017; Johnson et al., 2017; Antol et al., 2015; Suhr et al., 2017; Goyal et al., 2017; Suhr et al., 2018). Researchers develop models by joining features, aligning representation spaces, and using Transformers (Li et al., 2019b; Tan and Bansal, 2019). However, generalizability is still an issue when operating on unobserved data. It is hard for deep learning models to capture high-order patterns of reasoning, which is essential for their generalizability.

There are several challenging research directions for addressing learning representations for cross-modality data and enabling reasoning for target tasks. First is the alignment of the representation spaces for multiple modalities; second is designing architectures with the ability to capture high-order relations for generalizability of reasoning; third is using pre-trained modules to make the most use of minimal data.

An orthogonal direction to the above-mentioned aspects of learning is finding relevance between the components and the structure of various modalities when working with multi-modal data. Most of the previous language and visual reasoning models try to capture the relevance by learning representations based on an attention mechanism. Finding relevance, known as matching, is a fundamental task in information retrieval (IR) (Mitra et al., 2017). Benefiting from matching, Transformer models gain the excellent ability to index, retrieve, and combine features of underlying instances by a matching score (Vaswani et al., 2017), which leads to the state-of-the-art performance in various tasks (Devlin et al., 2019). However, the matching in the attention mechanism is used to learn a set of weights to highlight the importance of various components.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7642–7651
July 5 - 10, 2020.

In our proposed model, we learn representations directly based on the relevance score inspired by the ideas from IR models. In contrast to the attention mechanism and Transformer models, we claim that **the relevance patterns are as important**. With proper alignment of the representation spaces of different input modalities, matching can be applied to those spaces. The idea of learning relevance patterns is similar to Siamese networks (Koch et al., 2015) which learn transferable patterns of similarity of two image representations for one-shot image recognition. Similarity metric between two modalities is shown to be helpful for aligning multiple spaces of modalities (Frome et al., 2013).

The contributions of this work are as follows: **1)** We propose a cross-modality relevance (CMR) framework that considers entity relevance and high-order relational relevance between the two modalities with an alignment of representation spaces. The model can be trained end-to-end with customizable target tasks. **2)** We evaluate the methods and analyze the results on both VQA and NLVR tasks using VQA v2.0 and NLVR^2 datasets respectively. We improve state-of-the-art on both tasks' published results. Our analysis shows the significance of the patterns of relevance for the reasoning, and the CMR model trained on NLVR^2 boosts the training efficiency of the VQA task.

2 Related Work

Language and Vision Tasks. Learning and decision making based on natural language and visual information has attracted the attention of many researchers due to exposing many interesting research challenges to the AI community. Among many other efforts (Lin et al., 2014; Krishna et al., 2017; Johnson et al., 2017), Antol et al. proposed the VQA challenge that contains open-ended questions about images that require an understanding of and reasoning about language and visual components. Suhr et al. proposed the NLVR task that asks models to determine whether a sentence is true based on the image.

Attention Based Representation. Transformers are stacked self-attention models for general purpose sequence representation (Vaswani et al., 2017). They have been shown to achieve extraordinary success in natural language processing not only for better results but also for efficiency due to their parallel computations. Self-attention is a mechanism to reshape representations of components based on relevance scores. They have been shown to be effective in generating contextualized representations for text entities. More importantly, there are several efforts to pre-train huge Transformers based on large scale corpora (Devlin et al., 2019; Yang et al., 2019; Radford et al., 2019) on multiple popular tasks to enable exploiting them and performing other tasks with small corpora. Researchers also extended Transformers with both textual and visual modalities (Li et al., 2019b; Sun et al., 2019; Tan and Bansal, 2019; Su et al., 2020; Tsai et al., 2019). Sophisticated pre-training strategies were introduced to boost the performance (Tan and Bansal, 2019). However, as mentioned above, modeling relations between components is still a challenge for the approaches that try reshaping the entity representation space while the relevance score can be more expressive for these relations. In our CMR framework, we model high-order relations in relevance representation space rather than the entity representation space.

Matching Models. Matching is a fundamental task in information retrieval (IR). There are IR models that focus on comparing the global representation matching (Huang et al., 2013; Shen et al., 2014), the local components (*a.k.a* terms) matching (Guo et al., 2016; Pang et al., 2016), and hybrid methods (Mitra et al., 2017). Our relevance framework is partially inspired by the local components matching which we apply here to model the relevance of the components of the model's inputs. However, our work differs in several significant ways. First, we work under the cross-modality setting. Second, we extend the relevance to a high-order, *i.e.* model the relevance of entity relations. Third, our framework can work with different target tasks, and we show that the parameters trained on one task can boost the training of another.

3 Cross-Modality Relevance

Cross-Modality Relevance (CMR) aims to establish a framework for general purpose relevance in various tasks. As an end-to-end model, it encodes the relevance between the components of input modalities under task-specific supervision. We further add a high-order relevance between relations that occur in each modality.

Figure 1 shows the proposed architecture. We first encode data from different modalities with single modality Transformers and align the encoding

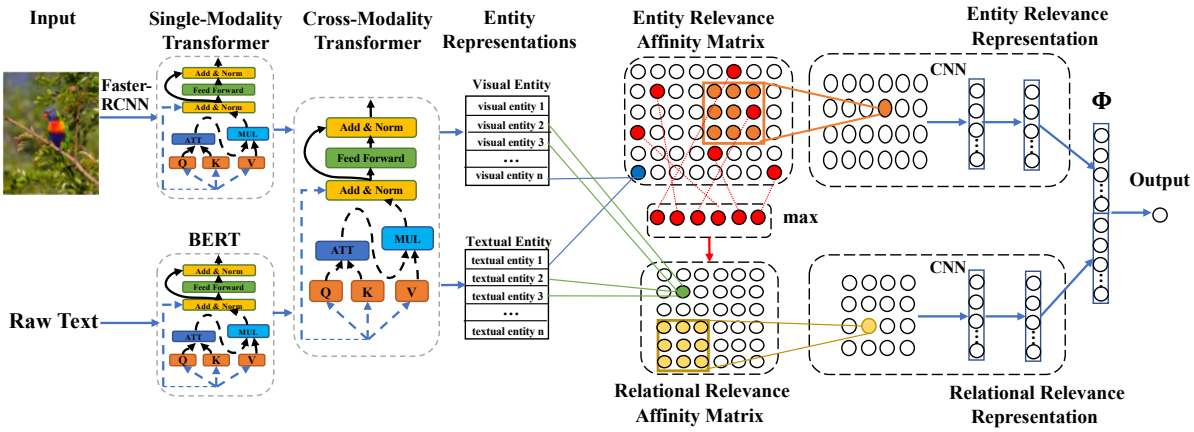

Figure 1: Cross-Modality Relevance model is composed of single-modality transformer, cross-modality transformer, entity relevance, and high-order relational relevance, followed by a task-specific classifier.

spaces by a cross-modality Transformer. We consistently refer to the words in text and objects in images (*i.e.* bounding boxes in images) as "entities" and their representations as "Entity Representations". We use the relevance between the components of the two modalities to model the relation between them. The relevance includes the relevance between their entities, as shown in the "Entity Relevance", and high-order relevance between their relations, as shown in the "Relational Relevance". We learn the representations of the affinity matrix of relevance score by convolutional layers and fully-connected layers. Finally, we predict the output by a non-linear mapping based on all the relevance representations. This architecture can help to solve tasks that need reasoning on two modalities based on their relevance. We argue that the parameters trained on one task can boost the training of the other tasks that deal with multi-modality reasoning.

In this section, we first formulate the problem. Then we describe our cross-modality relevance (CMR) model for solving the problem. The architecture, loss function, and training procedure of CMR are explained in detail. We will use the VQA and NLVR tasks as showcases.

3.1 Problem Formulation

Formally, the problem is to model a mapping from a cross-modality data sample $\mathcal{D} = \{\mathcal{D}_\mu\}$ to an output y in a target task, where μ denotes the type of modality. And $\mathcal{D}_\mu = \left\{d_1^\mu, \cdots, d_{N^\mu}^\mu\right\}$ is a set of entities in the modality μ. In visual question answering, VQA, the task is to predict an answer given two modalities, that is a textual question ($\mathcal{D}_t$) and a visual image ($\mathcal{D}_v$). In NLVR, given a textual statement ($\mathcal{D}_t$) and an image ($\mathcal{D}_v$), the task is to determine the correctness of the textual statement.

3.2 Representation Spaces Alignment

Single Modality Representations. For the textual modality $\mathcal{D}_t$, we utilize BERT (Devlin et al., 2019) as shown in the bottom-left part of Figure 1, which is a multi-layer Transformer (Vaswani et al., 2017) with three different inputs: WordPieces embeddings (Wu et al., 2016), segment embeddings, and position embeddings. We refer to all the words as the entities in the textual modality and use the BERT representations for textual single-modality representations $\left\{s_1^t, \cdots, s_{N^t}^t\right\}$. We assume to have N^t words as textual entities.

For visual modality $\mathcal{D}_v$, as shown in the top-left part of Figure 1, Faster-RCNN (Ren et al., 2015) is used to generate regions of interest (ROIs), extract dense encoding representations of the ROIs, and predict the probability of each ROI. We refer to the ROIs on images as the visual entities $\{d_1^v, \cdots, d_{N^v}^v\}$. We consider a fixed number, N^v, of visual entities with highest probabilities predicted by Faster-RCNN each time. The dense representation of each ROI is a local latent representation of a 2048-dimensional vector (Ren et al., 2015). To enrich the visual entity representation with the visual context, we further project the vectors with feed-forward layers and encode them by a single-modality Transformer as shown in the second column in Figure 1. The visual Transformer takes the dense representation, segment embedding, and pixel position embedding (Tan and Bansal, 2019) as input and generates the single-modality

representation $\{s_1^v, \cdots, s_{N^v}^v\}$. In case there are multiple images, for example, NLVR data (NLVR^2) has two images in each example, each image is encoded by the same procedure and we keep N^v visual entities per image. We refer to this as different sources of the same modality throughout the paper. We restrict all the single-modality representations to be vectors of the same dimension d. However, these original representation spaces should be aligned.

Cross-Modality Alignment. To align the single-modality representations in a uniformed representation space, we introduce a cross-modality Transformer as shown in the third column of Figure 1. All the entities are treated uniformly in the modality Transformer. Given the set of entity representations from all modalities we define the matrix with all the elements in the set $S = \left[s_1^t, \cdots, s_{N^t}^t, s_1^v, \cdots, s_{N^v}^v\right] \in \mathbf{R}^{d\times(N^t+N^v)}$. Each cross-modality self-attention calculation is computed as follows (Vaswani et al., 2017)[1],

$$\text{Attention}\left(K, Q, V\right) = \text{softmax}\left(\frac{K^\top Q}{\sqrt{d}}\right) V, \tag{1}$$

where in our case the key K, query Q, and value V, all are the same tensor S, and softmax $(\cdot)$ normalizes along the columns. A cross-modality Transformer layer consists of a cross-modality self-attention representation followed by residual connection with normalization from the input representation, a feed-forward layer, and another residual connection normalization. We stack several cross-modality Transformer layers to get a uniform representation over all modalities. We refer to the resulting uniformed representations as the entity representation and denote the set of the entity representations of all the entities as $\left\{s_1^{'t}, \cdots, s_{N^t}^{'v}, s_1^{'v}, \cdots, s_{N^v}^{'v}\right\}$. Although the representations are still organized by their original modalities per entity, they carry the information from the interactions with the other modality and are aligned in uniform representation space. The entity representations, as the fourth column in Figure 1, alleviate the gap between representations from different modalities, as we will show in the ablation studies, and allow them to be matched in the following steps.

[1]Please note here we keep the usual notation of the attention mechanism for this equation. The notations might have been overloaded in other parts of the paper.

3.3 Entity Relevance

Relevance plays a critical role in reasoning ability, which is required in many tasks such as information retrieval, question answering, intra- and inter-modality reasoning. Relevance patterns are independent from input representation space, and can have better generalizability to unobserved data. To consider the entity relevance between two modalities $\mathcal{D}_\mu$ and $\mathcal{D}_\nu$, the entity relevance representation is calculated as shown in Figure 1. Given entity representation matrices $S'^\mu = \left[s_1'^\mu, \cdots, s_{N^\mu}'^\mu\right] \in \mathbf{R}^{d\times N^\mu}$ and $S'^\nu = \left[s_1'^\nu, \cdots, s_{N^\nu}'^\nu\right] \in \mathbf{R}^{d\times N^\nu}$, the relevance representation is calculated by

$$A^{\mu,\nu} = \left(S'^\mu\right)^\top S'^\nu, \tag{2a}$$

$$\text{M}\left(\mathcal{D}_\mu, \mathcal{D}_\nu\right) = \text{CNN}_{\mathcal{D}_\mu,\mathcal{D}_\nu}\left(A^{\mu,\nu}\right), \tag{2b}$$

where $A^{\mu,\nu}$ is the affinity matrix of the two modalities as shown in the right side of Figure 1. $A_{ij}^{\mu,\nu}$ is the relevance score of ith entity in $\mathcal{D}_\mu$ and jth entity in $\mathcal{D}_\nu$. $\text{CNN}_{\mu,\nu}(\cdot)$ is a CNN, corresponding to the sixth column of Figure 1, which contains several convolutional layers and fully connected layers. Each convolutional layer is followed by a max-pooling layer. Fully connected layers finally map the flatten feature maps to d-dimensional vector. We refer to $\Phi_{\mathcal{D}_\mu,\mathcal{D}_\nu} = \text{M}\left(\mathcal{D}_\mu, \mathcal{D}_\nu\right)$ as the entity relevance representation between μ and ν.

We compute the relevance between different modalities. For the modalities considered in this work, when there are multiple images in the visual modality, we calculate the relevance representation between them too. In particular, for VQA dataset, the above setting results in one entity relevance representation: a textual-visual entity relevance $\Phi_{\mathcal{D}_t,\mathcal{D}_v}$. For NLVR^2 dataset, there are three entity relevance representations: two textual-visual entity relevance $\Phi_{\mathcal{D}_t,\mathcal{D}_{v_1}}$ and $\Phi_{\mathcal{D}_t,\mathcal{D}_{v_2}}$, and a visual-visual entity relevance $\Phi_{\mathcal{D}_{v_1},\mathcal{D}_{v_2}}$ between two images. Entity relevance representations will be flattened and joined with other features in the next layer of the network.

3.4 Relational Relevance

We also consider the relevance beyond entities, that is, the relevance of the entities' relations. This extension allows our CMR to capture higher-order relevance patterns. We consider pair-wise non-directional relations between entities in each modality and calculate the relevance of the rela-

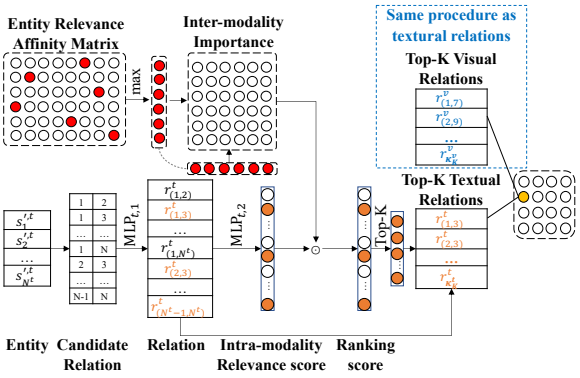

Figure 2: Relational Relevance is the relevance of top-K relations in terms of intra-modality relevance score and inter-modality importance.

tions across modalities. The procedure is similar to entity relevance as shown in Figure 1. We denote the relational representation as a non-linear mapping $\mathbf{R}^{2d} \rightarrow \mathbf{R}^{d}$ modeled by fully-connected layers from the concatenation of representations of the entities in the relation $r^{\mu}_{(i,j)} = \text{MLP}_{\mu,1}\left(\left[s^{'\mu}_{i}, s^{'\mu}_{j}\right]\right) \in \mathbf{R}^{d}$. Relational relevance affinity matrix can be calculated by matching the relational representation, $\left\{r^{\mu}_{(i,j)}, \forall i \neq j\right\}$, from different modalities. However, there will be $C^{2}_{N_\mu}$ possible pairs in each modality $\mathcal{D}_\mu$, most of which are irrelevant. The relational relevance representations will be sparse because of the irrelevant pairs on both sides. Computing the relevance score of all possible pairs will introduce a large number of unnecessary parameters which makes the training more difficult.

We propose to rank the relation candidates (i.e. pairs) by the intra-modality relevance score and the inter-modality importance. Then we compare the top-K ranked relation candidates between two modalities as shown in Figure 2. For the intra-modality relevance score, shown in the bottom left part of the figure, we estimate a normalized score based on the relational representation by a softmax layer.

$$U^{\mu}_{(i,j)} = \frac{\exp\left(\text{MLP}_{\mu,2}\left(r^{\mu}_{(i,j)}\right)\right)}{\sum_{k\neq l}\exp\left(\text{MLP}_{\mu,2}\left(r^{\mu}_{(k,l)}\right)\right)}. \quad (3)$$

To evaluate the inter-modality importance of a relation candidate, which is a pair of entities in the same modality, we first compute the relevance of each entity in text with respect to the visual objects. As shown in Figure 2, we project a vector that includes the most relevant visual object for each word, denoted this importance vector as v^{t}. This helps to focus on words that are grounded in the visual modality. We use the same procedure to compute the most relevant words to each visual object.

Then we calculate the relation candidates importance matrix V^{μ} by an outer product, $\otimes$, of the importance vectors as follows,

$$v^{\mu}_{i} = \max_{j} A^{\mu,\nu}_{ij}, \quad (4a)$$

$$V^{\mu} = v^{\mu} \otimes v^{\mu}, \quad (4b)$$

where v^{μ}_{i} is the ith scalar element in v^{μ} that corresponds to the ith entity, and $A^{\mu,\nu}$ is the affinity matrix calculated by Equation 2a.

Notice that the inter-modality importance V^{μ} is symmetric. The upper triangular part of V^{μ}, excluding the diagonal, indicates the importance of the corresponding elements with the same index in intra-modality relevance scores U^{μ}. The ranking score for the candidates is the combination (here the product) of the two scores $W^{\mu}_{(i,j)} = U^{\mu}_{(i,j)} \times V^{\mu}_{ij}$. We select the set of top-K ranked candidate relations $\mathcal{K}_\mu = \{\kappa_1, \kappa_2, \cdots, \kappa_K\}$. We reorganize the representation of the top-K relations as $R^{\mu} = [r^{\mu}_{\kappa_1}, \cdots r^{\mu}_{\kappa_K}] \in \mathbf{R}^{d\times K}$. The relational relevance representation between $\mathcal{K}_\mu$ and $\mathcal{K}_\nu$ can be calculated similar to the entity relevance representations as shown in Figure 1.

$$\text{M}\left(\mathcal{K}_\mu, \mathcal{K}_\nu\right) = \text{CNN}_{\mathcal{K}_\mu,\mathcal{K}_\nu}\left(\left(R^{\mu}\right)^{\top} R^{\nu}\right). \quad (5)$$

$\text{M}\left(\mathcal{K}_\mu, \mathcal{K}_\nu\right)$ has its own parameters which results in a d-dimensional feature space $\Phi_{\mathcal{K}_\mu,\mathcal{K}_\nu}$.

In particular, for VQA task, the above setting results in one relational relevance representation: a textual-visual relevance $\text{M}\left(\mathcal{K}_t, \mathcal{K}_v\right)$. For NLVR task, there are three entity relevance representations: two textual-visual relational relevance $\text{M}\left(\mathcal{K}_t, \mathcal{K}_{v_1}\right)$ and $\text{M}\left(\mathcal{K}_t, \mathcal{K}_{v_2}\right)$, and a visual-visual relational relevance $\text{M}\left(\mathcal{K}_{v_1}, \mathcal{K}_{v_2}\right)$ between two images. Relational relevance representations will be flattened and joined with other features in the next layers of the network.

After acquiring all the entity and relational relevance representations, namely $\Phi_{\mathcal{D}_\mu,\mathcal{D}_\nu}$ and $\Phi_{\mathcal{K}_\mu,\mathcal{K}_\nu}$, we concatenate them and use the result as the final feature $\Phi = \left[\Phi_{\mathcal{D}_\mu,\mathcal{D}_\nu}, \cdots, \Phi_{\mathcal{K}_\mu,\mathcal{K}_\nu}, \cdots\right]$. A task-specific classifier $\text{MLP}_{\Phi}\left(\Phi\right)$ predicts the output of the target task as shown in the right-most column in Figure 1.

3.5 Training

End-to-end Training. CMR can be considered as an end-to-end relevance representation extractor. We simply predict the output y from a specific task with the final feature Φ with a differentiable regression or classification function. The gradient of the loss function is back-propagated to all the components in CMR to penalize the prediction and adjust the parameters. We freeze the parameters of the basic feature extractors, namely BERT for textual modality and Faster-RCNN for visual modality. The parameters of the following parts will be updated by gradient descent: single modality Transformers (except BERT), the cross-modality Transformers, $\text{CNN}_{\mathcal{D}_\mu,\mathcal{D}_\nu}(\cdot)$, $\text{CNN}_{\mathcal{K}_\mu,\mathcal{K}_\nu}(\cdot)$, $\text{MLP}_{\mu,1}(\cdot)$, $\text{MLP}_{\mu,2}(\cdot)$ for all modalities and modality pairs, and the task-specific classifier $\text{MLP}_\Phi(\Phi)$.

The VQA task can be formulated as a multi-class classification that chooses a word to answer the question. We apply a softmax classifier on Φ and penalize with the cross-entropy loss. For NLVR^2 dataset, the task is binary classification that determines whether the statement is correct regarding the images. We apply a logistic regression on Φ and penalize with the cross-entropy loss.

Pre-training Strategy. To leverage the pre-trained parameters of our cross-modality Transformer and relevance representations, we use the following training settings. For all tasks, we freeze the parameters in BERT and faster-RCNN. We used pre-trained parameters in the (visual) single modality Transformers as proposed by (Tan and Bansal, 2019) and leave them being fine-tuned with the following procedure. Then we randomly initialize and train all the parameters in the model on NLVR with NLVR^2 dataset. After that, we keep and fine-tune all the parameters on the VQA task with the VQA v2.0 dataset. (See data description Section 4.1.) In this way, the parameters of the cross-modality Transformer and relevance representations, pre-trained by NLVR^2 dataset, are reused and fine-tuned on the VQA dataset. Only the final task-specific classifier with the input features Φ is initialized randomly. The pre-trained cross-modality Transformer and relevance representations help the model for VQA to converge faster and achieve a competitive performance compared to the state-of-the-art results.

4 Experiments and Results

4.1 Data Description

NLVR^2 (Suhr et al., 2018) is a dataset that aims to joint reasoning about natural language descriptions and related images. Given a textual statement and a pair of images, the task is to indicate whether the statement correctly describes the two images. NLVR^2 contains $107,292$ examples of sentences paired with visual images and designed to emphasize semantic diversity, compositionality, and visual reasoning challenges.

VQA v2.0 (Goyal et al., 2017) is an extended version of the VQA dataset. It contains $204,721$ images from the MS COCO (Lin et al., 2014), paired with $1,105,904$ free-form, open-ended natural language questions and answers. These questions are divided into four categories: Yes/No, Number, and Other.

4.2 Implementation Details

We implemented CMR using Pytorch[2]. We consider the 768-dimension single-modality representations. For textural modality, the pre-trained BERT "base" model (Devlin et al., 2019) is used to generate the single-modality representation. For visual modality, we use Faster-RCNN pre-trained by Anderson et al., followed by a five-layers Transformer. Parameters in BERT and Faster-RCNN are fixed. For each example, we keep 20 words as textual entities and 36 ROIs per image as visual entities. For the relational relevance, top-10 ranked pairs are used. For each relevance CNN, $\text{CNN}_{\mathcal{D}_\mu,\mathcal{D}_\nu}(\cdot)$ and $\text{CNN}_{\mathcal{K}_\mu,\mathcal{K}_\nu}(\cdot)$, we use two convolutional layers, each of which is followed by a max-pooling, and fully connected layers. For the relational representations and their intra-modality relevance score, $\text{MLP}_{\mu,1}(\cdot)$ and $\text{MLP}_{\mu,2}(\cdot)$, we use one hidden layer for each. The task-specific classifier $\text{MLP}_\Phi(\Phi)$ contains three hidden layers. The model is optimized using the Adam optimizer with $\alpha = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-6}$. The model is trained with a weight decay 0.01, a max gradient normalization 1.0, and a batch size of 32.

4.3 Baseline Description

VisualBERT (Li et al., 2019b) is an End-to-End model for language and vision tasks, consists of

[2]Our code and data is available at https://github.com/HLR/Cross_Modality_Relevance.

Models	Dev%	Test%
N2NMN	51.0	51.1
MAC-Network	50.8	51.4
FiLM	51.0	52.1
CNN+RNN	53.4	52.4
VisualBERT	67.4	67.0
LXMERT	74.9	74.5
CMR	**75.4**	**75.3**

Table 1: Accuracy on NLVR^2.

Model	Dev%	Test Standard%			
	Overall	Y/N	Num	Other	Overall
BUTD	65.32	81.82	44.21	56.05	65.67
ReGAT	70.27	86.08	54.42	60.33	70.58
ViLBERT	70.55	-	-	-	70.92
VisualBERT	70.80	-	-	-	71.00
BAN	71.4	87.22	54.37	62.45	71.84
VL-BERT	71.79	87.94	54.75	62.54	72.22
LXMERT	72.5	87.97	**54.94**	63.13	72.54
CMR	72.58	**88.14**	54.71	**63.16**	**72.60**

Table 2: Accuracy on VQA v2.0.

Transformer layers that align textual and visual representation spaces with self-attention. VisualBERT and CMR have a similar cross-modality alignment approach. However, VisualBERT only uses the Transformer representations while CMR uses the relevance representations.

LXMERT (Tan and Bansal, 2019) aims to learn cross-modality encoder representations from Transformers. It pre-trains the model with a set of tasks and fine-tunes on another set of specific tasks. LXMERT is the currently published state-of-the-art on both NLVR^2 and VQA v2.0.

4.4 Results

NLVR^2: The results of NLVR task are listed in Table 1. Transformer based models (VisualBERT, LXMERT, and CMR) outperform other models (N2NMN (Hu et al., 2017), MAC (Hudson and Manning, 2018), and FiLM (Perez et al., 2018)) by a large margin. This is due to the strong pre-trained single-modality representations and the Transformers' ability to reshape the representations that align the spaces. Furthermore, CMR shows the best performance compared to all Transformer-based baseline methods and achieves state-of-the-art. VisualBERT and CMR have similar cross-modality alignment approach. CMR outperforms VisualBERT by 12.4%. The gain mainly comes from entity relevance and relational relevance that model the relations.

VQA v2.0: In Table 2, we show the comparison with published models excluding the ensemble ones. Most competitive models are based on Transformers (ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019b), VL-BERT (Su et al., 2020), LXMERT (Tan and Bansal, 2019), and CMR). BUTD (Anderson et al., 2018; Teney et al., 2018), ReGAT (Li et al., 2019a), and BAN (Kim et al., 2018) also employ attention mechanism for a relation-aware model. The proposed CMR achieves the best test accuracy on Y/N questions and Other questions. However, CMR does not achieve the best performance on *Number* questions. This is because Number questions require the ability to count numbers in one modality while CMR focuses on modeling relations between modalities. Performance on counting might be improved by explicit modeling of quantity representations. CMR also achieves the best overall accuracy. In particular, we can see a 2.3% improvement over VisualBERT (Li et al., 2019b), as in the above mentioned NLVR^2 results. This shows the significance of the entity and relational relevance.

Another observation is that, if we train CMR for VQA task from scratch with random initialization while still use the fixed BERT and Faster-RCNN, the model converges after 20 epochs. As we initialize the parameters with the model trained on NLVR^2, it takes 6 epochs to converge. The significant improvement of convergence speed indicates that the optimal model for VQA is close to that of NLVR.

5 Analysis

5.1 Model Size

To investigate the influence of model sizes, we empirically evaluated CMR on NLVR^2 with various sets of Transformers sizes which contain the most parameters of the model. All other details are kept the same as descriptions in Section 4.2. Textual Transformer remains 12 layers because it is the pre-trained BERT. Our model contains $285M$ parameters. Among these parameters, around $230M$ parameters belong to pre-trained BERT and Transformer. Table 3 shows the results. As we increase the number of layers in the visual Transformer and the cross-modality Transformer, it tends to improve accuracy. However, the performance becomes stable when there are more than five layers. We choose five layers of visual Transformer and cross-modality Transformer in other experiments.

Textural	Visual	Cross	Dev%	Test%
12	3	3	74.1	74.4
12	4	4	74.9	74.7
12	5	5	75.4	**75.3**
12	6	6	75.5	75.1

Table 3: Accuracy on NLVR^2 of CMR with various Transformer sizes. The numbers in the left part of the table indicate the number of self-attention layers.

Models	Dev%	Test%
CMR	**75.4**	**75.3**
without Single-Modality Transformer	68.2	68.5
without Cross-Modality Transformer	59.7	59.1
without Entity Relevance	70.6	71.2
without Relational Relevance	73.0	73.4

Table 4: Test accuracy of different variations of CMR on NLVR^2.

5.2 Ablation Studies

To better understand the influence of each part in CMR, we perform the ablation study. Table 4 shows the performances of four variations on NLVR^2.

Effect of Single Modality Transformer. We remove both textual and visual single-modality Transformers and map the raw input with a linear transformation to d-dimensional space instead. Notice that the raw input of textual modality is the Word-Pieces (Wu et al., 2016) embeddings, segment embeddings, and the position embeddings of each word, while that of visual modality is the 2048-dimension dense representation of each ROI extracted by Faster-RCNN. It turns out that removing single-modality Transformers decreases the accuracy by 9.0%. Single modality Transformers play a critical role in producing a strong contextualized representation for each modality.

Effect of Cross-Modality Transformer. We remove the cross-modality Transformer and use single-modality representations as entity representations. As shown in Table 4, the model degenerates dramatically, and the accuracy decreases by 16.2%. The huge accuracy gap demonstrates the unparalleled contribution of the cross-modality Transformer to aligning representation spaces from input modalities.

Effect of Entity Relevance. We remove the entity relevance representation $\Phi_{\mathcal{D}_\mu,\mathcal{D}_\nu}$ from the final feature Φ. As shown in Table 4, the test accuracy is reduced by 5.4%. This is a significant difference of performance among Transformer based models (Li et al., 2019b; Lu et al., 2019; Tan and Bansal, 2019).

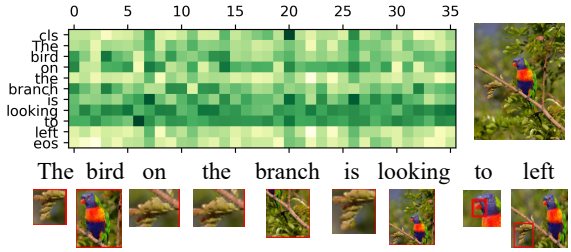

Figure 3: The entity affinity matrix between textual (rows) and visual (columns) modalities. The darker color indicates the higher relevance score. The ROIs with maximum relevance score for each word are shown paired with the words.

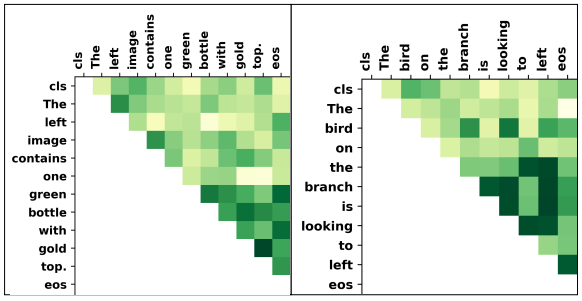

Figure 4: The relation ranking score of two example sentence. The darker color indicates the higher ranking score.

To highlight the significance of entity relevance, we visualize an example affinity matrix in Figure 3. The two major entities, "bird" and "branch", are matched perfectly. More interestingly, the three ROIs which are matching the phrase "looking to left" capture an indicator (the beak), a direction (left), and the semantic of the whole phrase.

Effect of Relational Relevance. We remove the entity relevance representation $\Phi_{\mathcal{K}_\mu,\mathcal{K}_\nu}$ from the final feature Φ. A 2.5% decrease in test accuracy is observed in Table 4. We argue that CMR models high-order relations, which are not captured in entity relevance, by modeling relational relevance. We present two examples of textual relation ranking scores in Figure 4. The learned ranking score highlights the important pairs, for example "gold - top", "looking - left", which describe the important relations in textual modality.

6 Conclusion

In this paper, we propose a novel cross-modality relevance (CMR) for language and vision reasoning. Particularly, we argue for the significance of relevance between the components of the two modalities for reasoning, which includes entity relevance

and relational relevance. We propose an end-to-end cross-modality relevance framework that is tailored for language and vision reasoning. We evaluate the proposed CMR on NLVR and VQA tasks. Our approach exceeds the state-of-the-art on NLVR^2 and VQA v2.0 datasets. Moreover, the model trained on NLVR^2 boosts the training of VQA v2.0 dataset. The experiments and the empirical analysis demonstrate CMR's capability of modeling relational relevance for reasoning and consequently its better generalizability to unobserved data. This result indicates the significance of relevance patterns. Our proposed architectural component for capturing relevance patterns can be used independently from the full CMR architecture and is potentially applicable for other multi-modal tasks.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This project is supported by National Science Foundation (NSF) CAREER award #1845771.

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 *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

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

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019a. Relation-aware graph attention network for visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10312–10321.

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. In *Advances in Neural Information Processing Systems*, pages 13–23.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1291–1299. International World Wide Web Conferences Steering Committee.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Iris D. Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. In *ACL*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472.

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

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232.

Yao-Hung Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.