Multimodal learning & Reasoning for Visual Q&A

-

- Sweta Kumari
- Neha Gupta
- Anshika Jain
- Nikita Garg

Visual Question & Answering

It is a task to provide answers to natural language questions about the contents of an image.

Does the man look happy?
What he is holding in his hand?

Multimodal learning & Reasoning for Visual Q&A

Solving the VQA task requires:

- Understanding image contents.
- Question words.
- Relationship between image content and question words.

The answering involves many computer vision tasks such as scene classification, object detection and classification, and face analysis.

Thus, it represents a problem comprised of multiple sub-problems over multimodal data.

Current VQA Models

- Some models are limited to a single task over single data modality.
- Image and text representations are joined via element-wise multiplication.
- Does not provide evidences of their reasoning capabilities.
- Completely remove a module that does not help in result.
- Models are Over-simplified deep neural network comprising of LSTM and CNN.

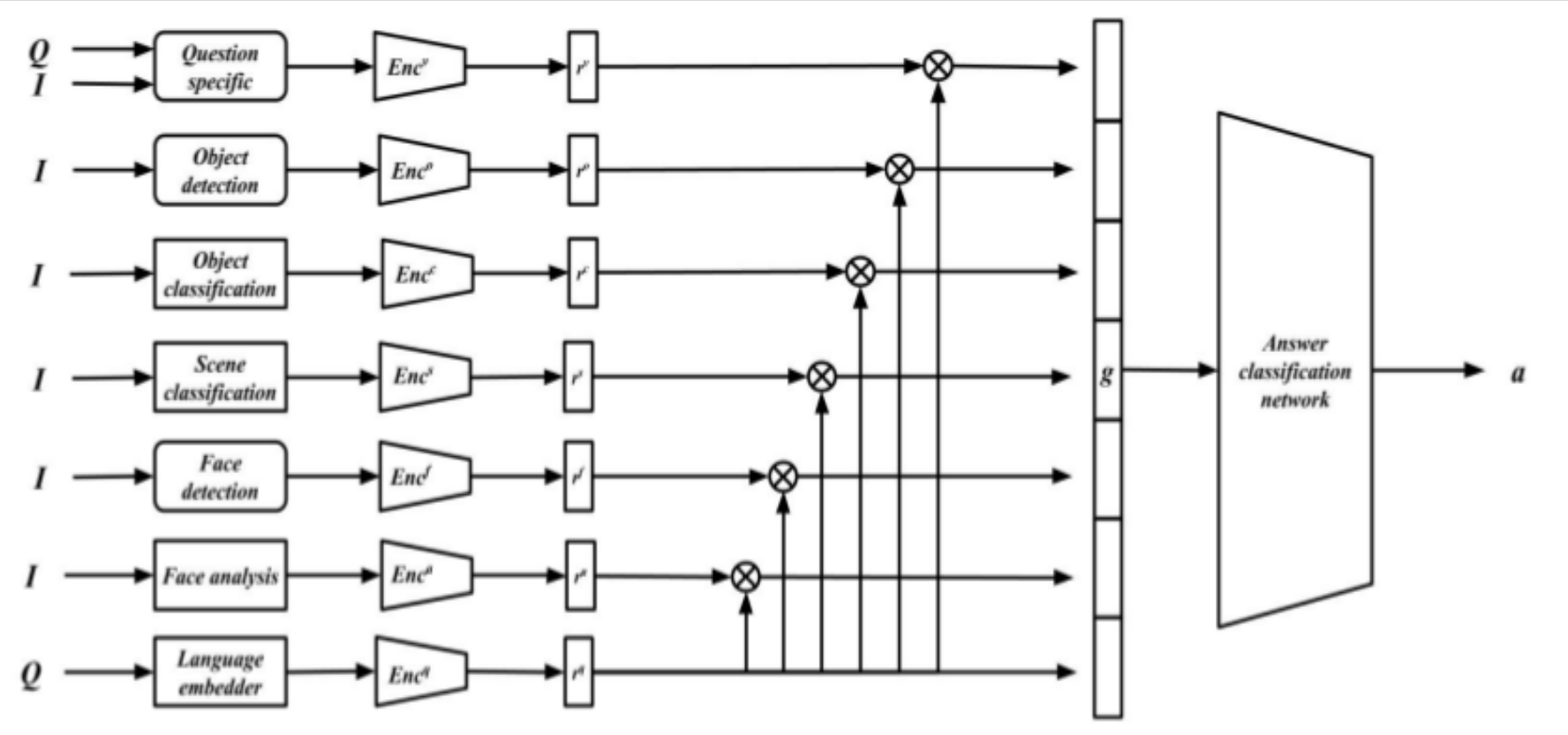
Proposed Model

- A model that learns to reason about entities in an image by learning a multifaceted representation of the image and the question.
- A novel multimodal representation learning and fusion method, crucial for obtaining the complete image understanding necessary for multimodal reasoning
- Perform an extensive evaluation and achieve new state-of-the-art performance on the two VQA benchmark dataset.

Model Used: ReasonNET

- **ReasonNet**, takes as input an image and a natural language text.
- Then, ReasonNet passes the image and the text through its different modules that encode multiple image aspects into multiple vector representations.
- At the same time, ReasonNet uses a language encoder module to encode the text into a vector representation.
- Finally, ReasonNet’s reasoning unit fuses the different representations into a single reasoning vector

Model Architecture



Representation Learning Modules:

Visual classification module(ϕ) : A visual classification module outputs a vector that contains the class probabilities of the specific image aspect assigned to that module.

Each visual classification module maps a different aspect of the image contents to a module specific class representation space.

$$c^{\phi} = \mathrm{vec}(P^{\phi} W^{\mathrm{LT}}),$$

where, $\mathbf{W}^{\mathbf{LT}}$ is the Lookup Table matrix

Visual Attention Module(φ):

ReasonNet passes the image I through a residual neural network to obtain a visual feature tensor $V \in R^{F \times W \times H}$ representing the global image contents.

Then it outputs a vector that contains visual features focused on the module's specific image aspect.

$$\boldsymbol{v}^{\varphi} = \sum_{i=1}^{W} \sum_{j=1}^{H} \boldsymbol{\alpha}_{i,j}^{\varphi} \boldsymbol{V}_{i,j}.$$

Contd...

Using the question representation vector $\mathbf{r}^l$ and a global visual feature tensor **V,**

ReasonNet learns a question-specific image representation.Attention probability distribution $\alpha^v \in \mathbf{R}^{W \times H}$ over the global visual feature tensor $\mathbf{V} \in \mathbf{R}^{f \times W \times H}$:

$$\boldsymbol{\alpha}^v = \text{softmax}\left[\boldsymbol{W}^\alpha \left(\sigma\left((\boldsymbol{W}^v \boldsymbol{r}^l + \boldsymbol{b}^v) \cdot \mathbb{1}\right) \circ \sigma(\boldsymbol{W}^V \boldsymbol{V} + \boldsymbol{b}^V)\right) + \boldsymbol{b}^\alpha\right],$$

where σ(·) := tanh(·), W ; $\boldsymbol{W}^v$; $\boldsymbol{W}^V$and the corresponding biases are learned parameters, $\mathbf{1} \in \mathbf{1}^{W \times H}$ is used to tile the question vector representation to match the V tensor dimensionality, and ◦ denotes element-wise matrix multiplication.

For the VQA task, ReasonNet incorporates the following set of modules:

1. Question-specific visual attention module
2. Object-specific visual attention module
3. Face-specific visual attention module
4. Object classification module
5. Scene classification module
6. Face analysis classification module.

Object Detection Module:

Given an image,as input the FCN outputs a set of object bounding boxes and their corresponding confidence scores. Each bounding box is represented as a four-element vector $\mathbf{d}^{\intercal}$ = [x, y, w, h]. Using a confidence score threshold ReasonNet obtains a set β containing high confidence bounding boxes. ReasonNet then uses the set β to compute an attention probability distribution $\boldsymbol{\alpha}^{o}$ that focuses the visual feature tensor **V** on the image objects.

$$\boldsymbol{\alpha}^{o} = \mathrm{softmax}(\hat{\boldsymbol{\alpha}}^{o}),$$

$$\hat{\alpha}^{o}_{i,j} = \max_{k=1,\ldots,|\mathcal{B}|} (\gamma^{k}_{i,j}),$$

$$\gamma^{k}_{i,j} = \begin{cases} 1 & d^{k}_{x} \leq i \leq (d^{k}_{x} + d^{k}_{w}), \\ & d^{k}_{y} \leq j \leq (d^{k}_{y} + d^{k}_{h}), \\ 0.1 & \text{otherwise.} \end{cases}$$

Object Classification Module:

ReasonNet crops out the image part corresponding to the box coordinates and then uses a residual network to classify the cropped-out image part and obtain a class label. The n class labels of the boxes with highest class probability are represented as n one-hot vectors of lookup table indices. The matrix $\mathbf{P}^{c}$, obtained by stacking the n vectors, is then mapped to a dense low-dimensional vector $\mathbf{r}^{c}$

Scene Classification Module:

ReasonNet uses a scene classification network as many of the questions explicitly or implicitly necessitate the knowledge of the image setting. The scene classification network is implemented as a residual network trained on the scene classification task. As before, the top n predicted class labels are represented as a matrix of n one-hot vectors $\mathbf{P}^s$ from which the module's representation vector $\mathbf{r}^s$ is obtained

Face Detection Module:

The face detector module is a fully convolutional network that outputs a set of face bounding boxes and confidence scores.

As with the object detector, ReasonNet uses a threshold to filter out bounding boxes with low confidence scores and obtain a set of face detections F. Then, from F,ReasonNet obtains an attention probability distribution $\mathbf{y}^f$ that focuses the visual feature tensor **V** on people's faces.

Face Analysis Module:

The face bounding boxes from F are also used to crop out the image regions that contain a face and using a convolutional neural network to obtain three class labels for each detected face representing the age group, the gender, and the emotion.

As with the other classification modules, ReasonNet represents the three class labels as a matrix of one-hot vectors $\mathbf{P}^a$ and obtain the face analysis representation vector $\mathbf{r}^a$

Encoder Units:

ReasonNet appends to each modules an encoder unit Enc that encodes a module's output vector **x** to a condensed vector **r** in a common low-dimensional representation space. The encoder units are implemented as two fully-connected layers followed by a non-linear activation function, and max pooling over the magnitude while preserving the sign:

$$\boldsymbol{r} = \mathrm{Enc}(\boldsymbol{x})$$

$$\mathrm{Enc}(\boldsymbol{x}) := \mathrm{sgn}(f(\boldsymbol{x})) \cdot \max(|f(\boldsymbol{x})|),$$

$$f(\boldsymbol{x}) := \sigma\big(\boldsymbol{W}_2^E\big(\sigma(\boldsymbol{W}_1^E\boldsymbol{x} + \boldsymbol{b}_1^E)\big) + \boldsymbol{b}_2^E\big),$$

Text Encoding

- The question words are converted to lowercase.
- All punctuation characters are removed.
- Uninformative words such as “a, “an”, “the”, etc. are removed.
- Questions are trimmed to contain atmost ten question words.
- The lookup table matrix is initialized with word2vec vectors.

ReasonNet treats the words in question as class labels and correspondingly uses the lookup table and an encoder unit to map the text to a vector $\mathbf{r}^l$ in a common low-dimensional representation space.

Bilinear Model

Learns the interaction between each module's representation $\mathbf{r}^k$ and the question representation $\mathbf{r}^l$.ReasonNet obtains a complete multimodal understanding by learning the interaction of the learned representations H = {r^v; r^o; r^c;r^s; r^f; r^a} with the question representation r^l

$$s_k = r_k^\top W_k^s r^l + b_k^s,$$

where k = 1,....,K and K is the number of representation learning modules, provides a rich vector representation s^k of the k-th module's output × language interaction.

Contd..

ReasonNet builds a complete image and language representation by concatenating each interaction vector s^k into a vector

$$g = \|_{k=1}^{K} s_k$$

where || denotes concatenation of vectors.

by observing which elements of the vector **g** are most active we can infer which modules ReasonNet used in the reasoning process and thus explain the reasons for its behavior.

Answer

The VQA problem is solved by modeling the likelihood probability distribution pvqa which for each answer a in the answer set Ω outputs the probability of being the correct answer, given a question Q about an image I:

$$\hat{a} = \arg\max_{a \in \Omega} p_{\text{vqa}}(a|Q, I; \boldsymbol{\theta}),$$

$$p_{\text{vqa}}(a|Q, I; \boldsymbol{\theta}) = \text{softmax}\left[\sigma\left(\boldsymbol{W}_2^g \sigma(\boldsymbol{W}_1^g \boldsymbol{g} + \boldsymbol{b}_1^g) + \boldsymbol{b}_2^g\right)\right],$$

where θ are the model parameters, â is the predicted answer, and Ω is the set of possible answers

Datasets Used:

The model is evaluated on the two benchmark VQA datasets:

- VQA v1.0
- VQA v2.0.

Test Result:

Table 1: Results of the ablation study on the VQA v2.0 validation.

Method	All	Y/N	Num	Other	Q-type changed
VQA	55.13	69.07	34.29	48.01	
VQA+Sc	56.80	70.62	35.14	49.99	+2.74% Which
VQA+Sc+oDec	58.46	71.05	36.16	52.86	+5.73% What color is the
VQA+Sc+oDec+oCls	59.82	72.88	37.38	54.47	+3.68% How
VQA+Sc+oDec+oCls+fDec	60.35	**74.21**	**37.46**	53.79	+12.63% Is the man
VQA+Sc+oDec+oCls+fDec+fAna	**60.60**	73.78	36.98	**54.81**	+0.88% Is he
ReasonNet-HadamardProduct	58.37	71.05	35.99	52.72	
ReasonNet-MCB [6]	58.78	71.04	36.96	53.35	
ReasonNet	**60.60**	**73.78**	**36.98**	**54.81**	

Our Implementation

Library Used :

PyTorch

Models Implemented :

ReasonNet model for Object Classification module.

Skip-gram model for word2vec

Tested on Dataset :

CIFAR - 10 dataset

THANK YOU