Visual Question Answering

Farooqui Nabeel Ahmed, Dattrey Khansole, Vikash Kumar
Under the guidance of
Dr. CHANDRAMANI CHAUDHARY
Assistant Professor - CSED
National Institute of Technology Calicut

***Abstract*—Visual Question Answering (VQA) aims to enable machines to comprehend and respond to questions posed about images. The core objective of Visual Question Answering is to develop algorithms that can understand textual questions with respect to visual cues present in the images, and then generate accurate textual answers. VQA is all about making the computer smart enough to understand the picture and give the right answer. This task involves both textual and visual information. Each of these modalities has made tremendous progress in recent couple of years. We will be exploring models with respect to each individual modality and the models which use both visual and textual modalities.**

I. Introduction

The main objective of Visual Question answering is Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. As questions can be multiple choice format or one word answer. Our model will be able to detect the type of question and generate answers accordingly on its own.

II. Methodology

In this section, we outline the methodology of our project,detailing the sequential stages that lead to the implementation of a Visual Question Answering(VQA).
Given a pair of image and sentence, Model takes the visual regions of the image and textual tokens of the sentence as inputs. We design an Image Embedder and a Text Embedder to extract their respective embeddings. These embeddings are then fed into a multi-layer Transformer to learn a cross-modality contextualized embedding across visual regions and textual tokens.

- **Image Embedder:**
  Initially, we employ Faster R-CNN to extract visual features, specifically pooled Region of Interest (ROI) features, for each region of interest. Additionally, we encode the location features for each region using a 7-dimensional vector. Subsequently, both the visual and location features undergo processing through a fully-connected (FC) layer, projecting them into a unified embedding space. The ultimate visual embedding for each region is derived by summing the outputs of the two FC layers and then passing the result through a layer normalization (LN) layer.
- **Text Embedder:**
  Text Embedder, we follow BERT and tokenize the input sentence into WordPieces . The final representation for each sub-word token is obtained via summing up its word embedding and position embedding, followed by another LN layer.
- **Transformer**: a Transformer module is applied to learn generalizable contextualized embeddings for each region and each word through well-designed pre-training task

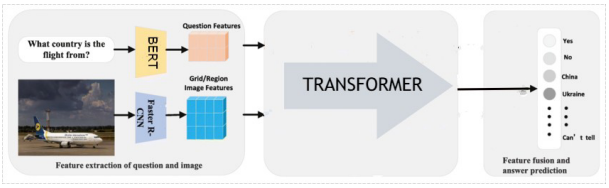

Fig. 1. FlowChart

III. Pre-Training Task

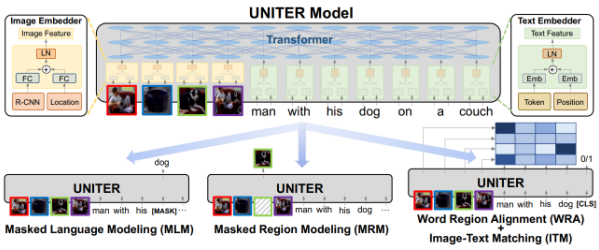

Fig. 2. Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer Transformer, learned through four pre-training tasks

- **Masked Language Modeling (MLM):**
  We denote the image regions as $v = \{v_1, \ldots, v_K\}$, the input words as $w = \{w_1, \ldots, w_T\}$, and the mask indices as $m \in N^M$. In MLM, we randomly mask out the input words with a probability of 15% and replace the masked ones W_m with the special token [MASK]. The goal is to predict these masked words based on the observation of

their surrounding words $w_{\setminus m}$ and all image regions v, by minimizing the negative log-likelihood:

$$L_{MLM}(\theta) = -E(w,v) \sim_D \log P_\theta(w_m|w_{\setminus m}, v) \quad (1)$$

where θ is the trainable parameters. Each pair (w, v) is sampled from the whole training set D

- **Image-Text Matching (ITM):**
In ITM, a special token [CLS] represents the fused modalities. The model takes a sentence and image regions, producing a binary label $y \in \{0, 1\}$ indicating a match. The [CLS] token represents the joint image-text pair, processed through an FC layer and sigmoid function to predict a score $s_\theta(w, v) \in [0, 1]$. Training involves sampling positive/negative pairs (w, v) from dataset D, with negative pairs created by swapping image or text. Binary cross-entropy loss is applied for optimization.
- **Word-Region Alignment (WRA) :** Word-Region Alignment (WRA) utilizes Optimal Transport (OT) for alignment, where a transport plan $T \in R^{T\times K}$ is learned to optimize the alignment between w and v. OT exhibits characteristics suitable for WRA: (i) Self-normalization, where all elements of T sum to 1. (ii) Sparsity, as OT yields a sparse solution T with at most $(2r - 1)$ non-zero elements, where $r = \max(K, T)$, leading to a more interpretable and robust alignment . (iii) Efficiency, as our solution, can be readily obtained using iterative procedures requiring only matrix-vector products , making it applicable to large-scale model pre-training.
Specifically, (w, v) can be viewed as two discrete distributions μ, ν, formulated as $\mu = \sum_{i=1}^{T} a_i \delta_{w_i}$ and $\nu = \sum_{j=1}^{K} b_j \delta_{v_j}$, with δ_{w_i} as the Dirac function centered on w_i. The weight vectors $a = \{a_i\}_{i=1}^{T} \in \Delta^T$ and $b = \{b_j\}_{j=1}^{K} \in \Delta^K$ belong to the T- and K-dimensional simplex, respectively (i.e., $\sum_{i=1}^{T} a_i = \sum_{j=1}^{K} b_j = 1$), as both μ and ν are probability distributions. The OT distance between μ and ν (thus also the alignment loss for the (w, v) pair) is defined as:

$$L_{WRA}(\theta) = D_{ot}(\mu, \nu) = \min_{T\in\Pi(a,b)} \sum_{i=1}^{T}\sum_{j=1}^{K} T_{ij} \cdot c(w_i, v_j) \quad (2)$$

where $\Pi(a, b) = \{T \in R_+^{T\times K} | T1_m = a, T^T 1_n = b\}$, 1_n denotes an n-dimensional all-one vector, and $c(w_i, v_j)$ is the cost function evaluating the distance between w_i and v_j. In experiments, the cosine distance $c(w_i, v_j) = 1 - \frac{w_i^T v_j}{\|w_i\|_2 \|v_j\|_2}$ is used. The matrix T is denoted as the transport plan, interpreting the alignment between two modalities. Unfortunately, the exact minimization over T is computationally intractable, and we consider the IPOT algorithm to approximate the OT distance. After solving T, the OT distance serves as the WRA loss that can be used to update the parameters θ.
- **Masked Region Modeling (MRM):**
Similar to MLM, we sample image regions and mask their visual features with a probability of 15%. The model is trained to reconstruct the masked regions v_m given the remaining regions $v_{\setminus m}$ and all the words w. The visual features of the masked region are replaced by zeros. Unlike textual tokens represented as discrete labels, visual features are high-dimensional and continuous, making supervision via class likelihood impractical. Instead, we propose three variants for MRM, sharing the same objective base:

$$L_{MRM}(\theta) = E(w, v) \sim Df_\theta(v_m|v_{\setminus m}, w). \quad (3)$$

IV. Conclusion

In conclusion we studied all relevant datasets which are available for VQA. We also covered available deep learning architecture/models for VQA. Till now we have covered different types of dataset which include given an image and an open-ended natural language question about the image, the task is to provide an accurate natural language answer. Based on all dataset and deep learning models we have an initial design which include fundamental algorithm R-CNN for image feature extraction and BERT for processing given questions.

References

[1] Y.-C. Chen et al. "UNITER: UNiversal Image-TExt Representation Learning." In: AAAI (2020)..

[2] Kai Wang; Yun Pan; Xiang Yao 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP).

[3] Xiangyu Wu; Jianfeng Lu; Zhuanfeng Li; Fengchao Xiong 2022 IEEE International Conference on Image Processing (ICIP).

[4] Deepika Kanakamedala; Tilekya Veeranki; Reshitha Bitla; Sreeja Vangalapudi; Subetha. T 2021 Innovations in Power and Advanced Computing Technologies (i-PACT).

[5] Zichao Yang; Xiaodong He; Jianfeng Gao; Li Deng; Alex Smola 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," Adv. Neural Inf. Process. Syst,vol.2, no January, pp. 1682–1690, 2014..