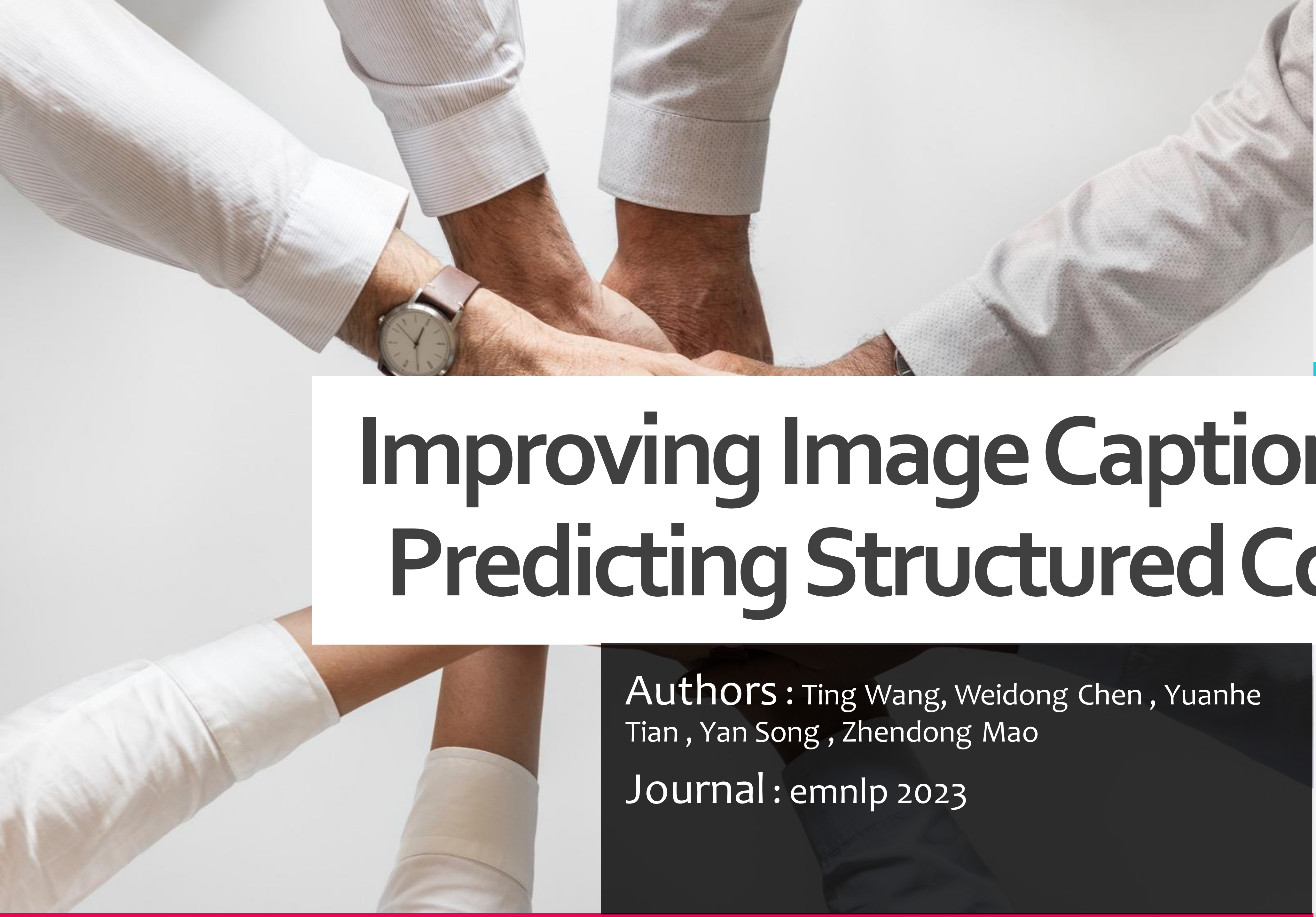


Improving Image Captioning via Predicting Structured Concepts

Authors : Ting Wang, Weidong Chen , Yuanhe Tian , Yan Song , Zhendong Mao

Journal : emnlp 2023

DL Project

Abstract :

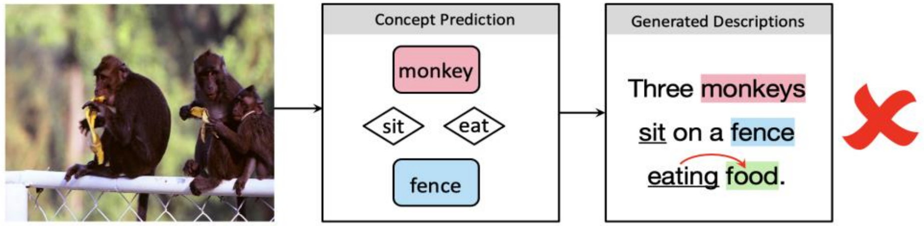
- Difficulty in solving the semantic gap between images and texts for image captioning.
- Conventional studies focused on using semantic concepts as a bridge between modalities, improving captioning performance.
- Previous studies often ignored relationships among concepts, crucial for both objects in images and word dependencies in text.
- Proposal of a Structured Concept Predictor (SCP) to predict concepts and their structures, integrated into captioning to enhance visual signal contribution
- Design of Weighted Graph Convolutional Networks (W-GCN) to depict concept relations driven by word dependencies, improving description generation by distinguishing cross-modal semantics.

Introduction :

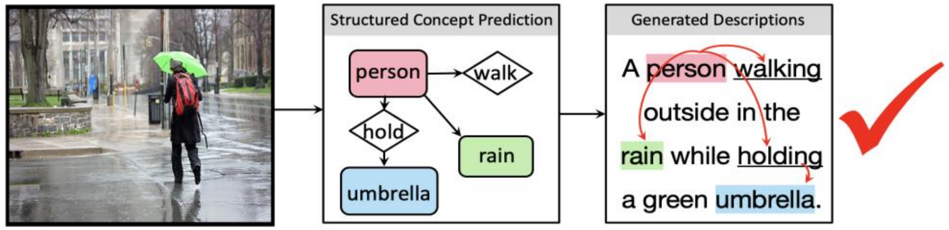
The image captioning task aims at generating a human-like description for a given image, normally requiring recognition and understanding of the content in the image, including objects, attributes, and their relationships, etc. The task is regarded as an interdisciplinary research of computer vision and natural language processing and has become a popular topic in recent years.

Current methods usually follow an encoder-decoder framework, using a pre-trained object detector/classifier as an encoder to mine visual information in an image, and then feeding it into an RNN or Transformer based decoder for description prediction along with partially generated words. However, in most cases, the extracted visual information is insufficient, even with the use of powerful visual feature extractors. This shortcoming makes the decoder rely too much on partially generated words to predict the remaining words to ensure the fluency of the generated description, that is, the model relies too much on linguistic priors during decoding, and sometimes the resulted words does not related to the image at all. In short, the major challenge that the image captioning facing now is that description generation relies too much on linguistic priors and has little to do with images.

- Illustrations of our motivation. Compared with integrating semantic concepts into the image captioning framework, we find that the structured concepts helps reduce over-reliance on linguistic priors in language generation.



- it is seen that concepts have relations, which are not only shown among objects in the image, but also in word dependencies in the text. Structured semantic concepts improve the performance of the model from the perspective of language generation.


Structured Concept Prediction
person
walk
hold
rain
umbrella
Generated Descriptions
A person walking outside in the rain while holding a green umbrella.

- we propose a structured concept predictor (SCP) to improve image captioning, which not only integrates concept prediction into the end-to-end image captioning, but also predicts the structures of obtained concepts based on the word dependencies. Specifically, we propose weighted graph convolutional networks (W-GCN), with its input graph built based on mutual information priors of all descriptions in an unsupervised manner. The mutual information priors are the probability of co-occurrence of two words within a certain distance in the same description, making the language generation no longer limited to local contextual information and measuring the relationship between two concept words well.

Problem Definition:

Image captioning needs balance between linguistic priors & visual info. Current methods often rely too much on linguistic priors, disconnecting from images. Challenge: reliance on linguistic priors hampers use of visual info, leading to descriptions unrelated to images. Existing methods use encoder-decoder frameworks with pre-trained object detectors but face insufficiency in visual features. Augmenting visual info or emphasizing semantic concepts helps but overlooks concept relationships, exacerbating overfitting on linguistic priors. Structured semantic concepts offer potential to enhance language generation. Need: integrate concept prediction, predict concept structures based on word dependencies. Mitigate over-reliance on linguistic priors by incorporating structured semantic concepts into image captioning. Leverage techniques like weighted graph convolutional networks to learn concept relationships, reducing linguistic priors during captioning, ensuring description generation remains closely associated with image content.

Description of the Method :

- **Visual Feature Processing Extractor :** To generate descriptions, the first step is to extract the visual features from images. Following (Li et al., 2022b), in order to narrow the semantic gap between images and text, and help semantic alignment, we extract visual features from images I by using the encoder of CLIP with the ResNet-101 (He et al., 2016) backbone, which has the ability to understand complex scenarios, and having excellent domain generalization ability after pre-training with a large dataset. The process can be formulated as:

$$X = f_v(I) \qquad (1)$$

where fv is the visual extractor, and X is the image features.

- **Encoder :** Since the image feature is in the form of 2D, we first flatten X into a sequence {x1, x2, ..., xS}, xs ∈ R d , where xs are patch features and d is the size of the feature vector. Then we employ Nv Transformer encoder blocks to further encode image features as a sequence. Outputs are the hidden states encoded from the input features X extracted from the visual extractor. The whole process can be formulized as follows:

$$H_v^{(0)} = X \qquad (2)$$

$$\overline{H}_v^{(l)} = \mathrm{MHA}(H_v^{(l-1)}, H_v^{(l-1)}, H_v^{(l-1)}) \qquad (3)$$

$$H_v^{(l)} = \mathrm{LN}(\overline{H}_v^{(l)} + H_c^{(l-1)}) \qquad (4)$$

- **Concept Prediction :** Specifically, we predict the semantic concepts under the guidance of visual features through a set of concept queries Q, by leveraging Nc Transformer encoder blocks. The set of learnable queries learns the essential concepts within the images. Through the image interaction, each of our learnable queries focuses on a specific area of the image and learns the information (concepts) contained in the image. These concepts include objects, relative positions between objects, actions, etc. Each Transformer block reinforces concept queries by interacting visual features Ve with object queries Q through a cross-attention mechanism. The whole process can be formulized as follows:

$$H_c^{(0)} = Q \qquad (5)$$

$$H_c^{(l)} = \mathrm{LN}(\mathrm{MHA}(H_c^{(l-1)}, \widetilde{V}, \widetilde{V}) + H_c^{(l-1)}) \quad (6)$$

where H (l) c indicates the output of the l-th middle hidden layer and the superscript indicates the number of layer. Thus, the output of the Transformer encoder HNc c is the output of the last Transformer layer.

We then feed the output of the last block into a multi-linear perception network to get concept features C:

$$C = \mathrm{MLP}(H_c^{N_c}) \qquad (7)$$

where MLP is the multi-linear perception network with the sigmoid activation.

During training, following (Fang et al., 2022), we describe it as a multi-label classification problem. Due to the imbalance of the distribution of concepts, we use asymmetric loss (Ben-Baruch et al., 2020), which can handle the sample imbalance problem of multi-label classification tasks 362 well.

Asymmetric loss is calculated for concept prediction:

$$\mathcal{L}_c = \mathbf{asym}(C, Y_c) \qquad (8)$$

where Yc denote the visual concept of the groundtruth sentence that corresponds to the concept vocabulary

Weighted Graph Convolutional Network : we propose to construct graph for these concepts, explore the relationship between them by Weighted Graph Convolutional Networks (W-GCN), and obtain structured concepts. Structure concepts learns to estimate the linguistic relative position of semantic word pairs, thereby allocating all the semantic words in potential linguistic order as humans. In doing so, the output sequence of structured semantic concepts serve as additional visually-grounded language priors, which encourage the visual contribution in generation. Concretely, the nodes of the graph represent concepts G = {g1, g2, ..., gk}, and the edges represent the relationship between nodes gi and gj for ∀i, j ∈ {1, 2, ..., k}, which can be represented by an adjacency matrix A. In A, aij = 1, if there is an edge between gi and gj or i = j, otherwise aij = 0.

Graph Construction : As the obtained semantic concepts cannot form a complete sentence, our method cannot leverage existing dependency parsers to estimate their relations. Without such a parser, we need an alternative way to find satisfied word pairs to build initial graphs in our W-GCN, which equivalent to build the initial adjacency matrix A. Inspired by the studies (Tian et al., 2020) which leverage chunks (n-grams) as additional features to carry contextual information, we propose to construct the graph based on the word dependencies extracted from a pre-constructed n-gram lexicon D. Specifically, we count the frequency of the occurrences of each word and the frequency of simultaneous occurrence of any two words within NL word distance (considering the order) in all sentences of the training set. We regard two words within NL distance as they have word dependency. Then we calculate the Pointwise Mutual Information (PMI) score of any two words w1, w2 by the following formula and set a threshold to determine if they are strongly correlated.

$$PMI(w_1, w_2) = log\frac{p(w_1 w_2)}{p(w_1)p(w_2)} \qquad (9)$$

where p(w1), p(w2)is the probability of w1, w2 in the training set, p(w1, w2) is the probability that both w1 and w2 are within NL word distance. We store all strongly correlated word pairs in a word lexicon D and refer to it to build the graph. If the concept represented by the two nodes in the graph can be found in the lexicon D, then the corresponding element of its adjacency matrix is initialized as correlated, which is set to 1. Otherwise, setting the value to 0.

The Weighted GCN : Based on the adjacency matrix, the W-GCN module of the L layers can learn from all the input concepts. Considering that the contribution of different gj to gi may be different, we further apply the attention mechanism to the adjacency matrix, replacing aij with the weights αij . For each gi and all its related gj , we calculate weight αij for the concept pair. In particular, at the l-th layer, for each gi , all the gj associated with it can be calculated:

$$\alpha_{ij}^{(l)} = \frac{a_{ij} \cdot exp(h_i^{(l-1)} \cdot W_{pos}^{(l)} \cdot h_j^{(l-1)})}{\sum_{j=1}^{n} a_{ij} \cdot exp(h_i^{(l-1)} \cdot W_{pos}^{(l)} \cdot h_j^{(l-1)})} \qquad (10)$$

$$h_i^{(l)} = \sigma(\mathrm{LN}(\sum_{j=1}^{n} \alpha_{ij}(W^{(l)} \cdot h_j^{(l-1)} + b^{(l)}))) \quad (11)$$

where W (l) pos is a trainable parameter, it can model the position relationship between gi and gj (three choices: W (l) lef t, W(l) right, W(l) self). h (l−1) i is the hidden vector from layer l – 1, W(l) and b (l) are the trainable matrices and biases of the W-GCN at layer l, LN is layer normalization, and σ is the ReLU activation function.

Language Decoder : The sentence decoder takes each word as input and learns to predict the next word auto-regressively conditioned on Ve and Ce. The formulations are

$$H_i = \mathrm{MHA}(w_i, \widetilde{V}, \widetilde{V}) + \mathrm{MHA}(w_i, \widetilde{C}, \widetilde{C}) \quad (12)$$

$$y_{i+1} = \mathrm{LN}(H_i + w_i) \qquad (13)$$

$$\mathcal{L}_{cap} = \sum_{t=1}^{T} \mathrm{CE}(Y, Y_{gt})$$

$$\mathcal{L} = \mathcal{L}_{cap} + \beta \cdot \mathcal{L}_{c}. \qquad (15)$$

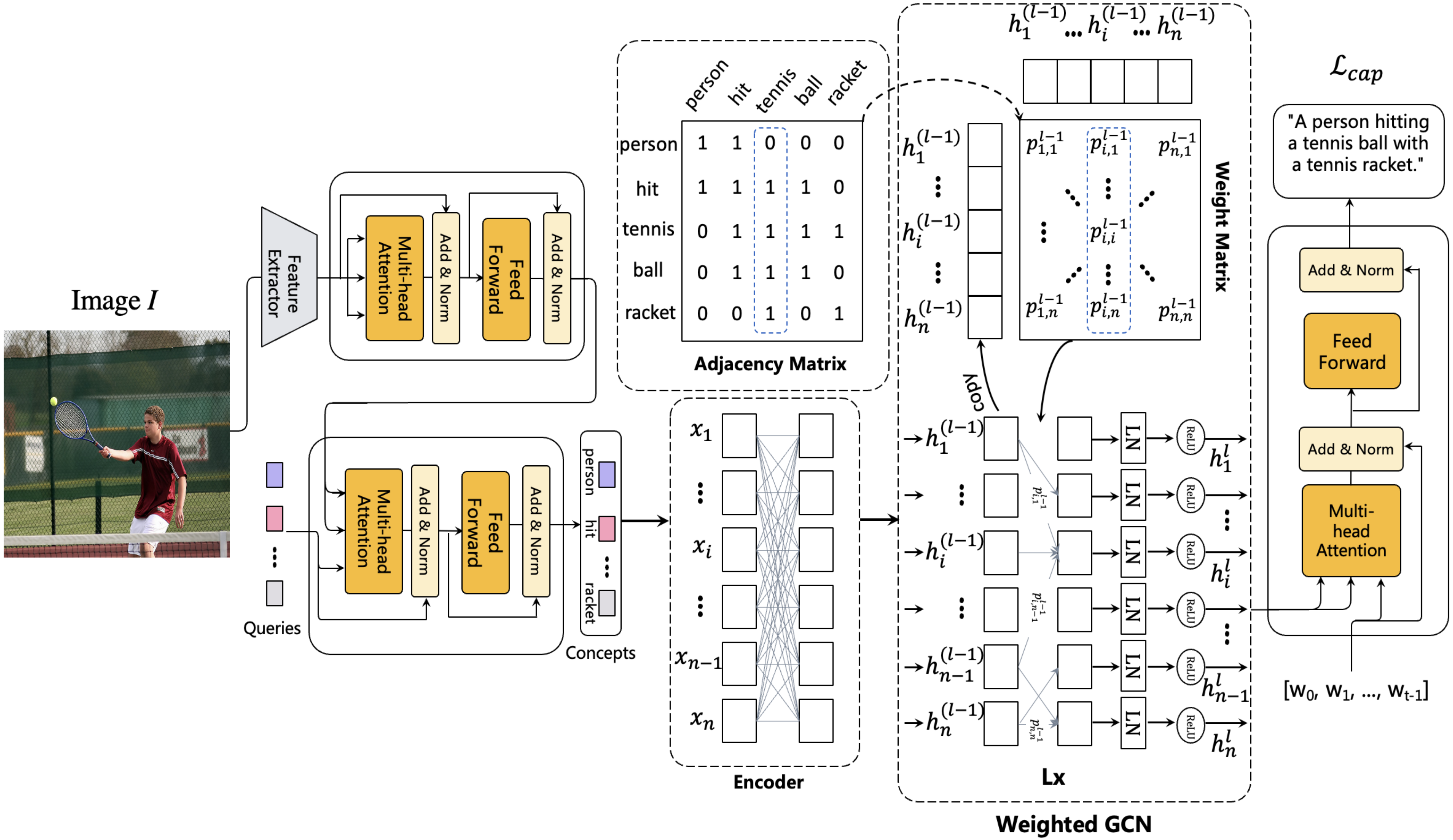
where yi+1 is the (i + 1)th word of the predicted sentence, and Hi is the hidden state. Y = [y1, y2, ..., yT] is the predicted sentence. the total loss is the combination of visual concept prediction loss and the language prediction loss

Implementation :

- **Datasets and Metrics :** Our experiments are conducted on the MS COCO (Lin et al., 2014), which is the most popular image captioning benchmark dataset. It consists of more than 120,000 images, and each image is equipped with five human-annotated descriptions. We follow Karpathy's split, which divides 5,000 images for validation, 5,000 images for testing, and the rest for training (Karpathy and Fei-Fei, 2015). For fair comparison with other techniques, we leverage pycocoevalcap package to calculate five evaluation metrics: BLEU-N (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016).

- **Implementation Details :** Our feature extractor is CLIP (Mokady et al., 2021) with the ResNet-101 (He et al., 2016) backbone and the dimension of the grid visual feature is 2048. Following previous work (Li et al., 2022b), to build concept vocabulary, we filter out low-frequency words and convert all uppercase letters to lowercase letters to all caption descriptions. Thus, a concept vocabulary that containing 906 words is constructed. The word distance in the pre-constructed lexicon D is 3. The number of layers in Weighted GCN is set to 2. The Transformer block in the Feature Encoder Module, the Concept Prediction Module and the Language Prediction Module are 3 layers, 6 layers, and 6 layers, respectively. The size of the hidden state features is set to 512. The query size is set to 17. β is set to 1 in this work. Our code is developed based on the COS-Net1 . The model is trained using a typical two-stage training method. In the first stage, we utilize Adam optimizer and cross-entropy loss with a learning rate of 0.0005 and take about one hour per epoch. In the second stage, the self-critical sequence training strategy is used to further optimize the CIDEr scoring model, and the learning rate is set to 0.00005, which takes about 4 hours per epoch. In inference, the beam size is set to 3. The number of parameters our model used is 20M. All experiments are conducted on a single RTX 3090.

Summary :



	person	hit	tennis	ball	racket
person	1	1	0	0	0
hit	1	1	1	1	0
tennis	0	1	1	1	1
ball	0	1	1	1	0
racket	0	0	1	0	1

Adjacency Matrix

Summary :

- We will be using the dataset MS COCO and we'll use Karpathy's split.
- We will start with CLIP (Mokady et al., 2021) with the ResNet-101 (He et al., 2016) backbone to extract the features of Images.
- Then we have to construct Encoder block to get more information of features.
- After that we have to form Adjacency matrix for words of feature.
- Then go for Weighted Graph Convolutional Networks.
- Then using SCP we'll generate Structures sentence with more concepts using Decoder block.

Thank You

Harsh Raj 👤

+91 7480034563 📱

harsh.raj@iitg.ac.in ✉