Learning a Recurrent Visual Representation for Image Caption Generation

Zhang, Liqiang

June 2, 2018

Abstract

In this paper, the author explore the bi-directional mapping between images and their sentence-based descriptions. They propose learning this mapping using a recurrent neural network. Using the same model, it can also reconstruct the visual features associated with an image given its visual description. They use a novel recurrent visual memory that automatically learns to remember long-term visual concepts to aid in both sentence generation and visual feature reconstruction. These include sentence generation, sentence retrieval and image retrieval. Results are better than or comparable to state-of-the-art results on the image and sentence retrieval tasks for methods using similar visual features.

1. Introduction

A good image description is often said to "paint a picture in your minds eye." The creation of a mental image may play a significant role in sentence comprehension in humans. In fact, it is often this mental image that is remembered long after the exact sentence is forgotten. In fact, it is often this mental image that is remembered long after the exact sentence is forgotten.

Recently, several papers have explored learning joint feature spaces for images and their descriptions. These approaches project image features and sentence features into a common space, which may be used for image search or for ranking image captions. Various approaches were used to learn the projection, including Kernel Canonical Correlation Analysis (KCCA), recursive neural networks, or deep neural networks. While these approaches project both semantics and visual features to a common embedding, they are not able to perform the inverse projection. That is, they cannot generate novel sentences or visual depictions from the embedding.

In this paper, the author propose a bi-directional representation capable of generating both novel descriptions from images and visual representations from descriptions. Critical to both of these tasks is a novel representation that dynamically captures the visual aspects of the scene that have already been described. That is, as a word is generated or read the visual representation is updated to reflect the new information contained in the word. the author accomplish this using Recurrent Neural Networks (RNNs) [1]. During sentence generation, these novel dynamically updated visual representation acts as a long-term memory of the concepts that have already been mentioned. This allows the network to automatically pick salient concepts to convey that have yet to be spoken.

2. Approach

First, this approach is to be able to generate sentences given a set of visual observations or features. Specifically, we want to compute the probability of a word w_t being generated at time t given the set of previously generated words $W_{t-1} = w_1, ..., w_{t-1}$ and the observed visual features V. Second, this approach is to enable the capability of computing the likelihood of the visual features V given a set of spoken or read words W_t for generating visual representations of the scene or for performing image search. To accomplish both of these tasks the author introduce a set of latent variables U_{t-1} that encodes the visual interpretation of the previously generated or read words W_{t-1}.

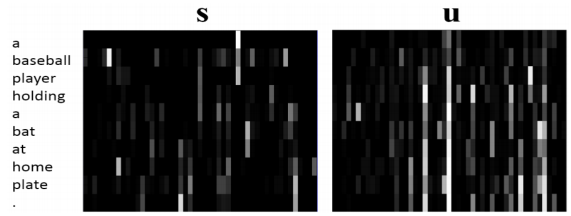

Figure 1. Illustration of the hidden units **s** and **u** activations through time (vertical axis). Notice that the visual hidden units **u** exhibit long-term memory through the temporal stability of some units, where the hidden units **s** change significantly each time step.

The goal of Using U is to compute $P(w_t|V, W_{t-1}, U_{t-1})$ and $P(V|W_{t-1}, U_{t-1})$. Combining these two likelihoods

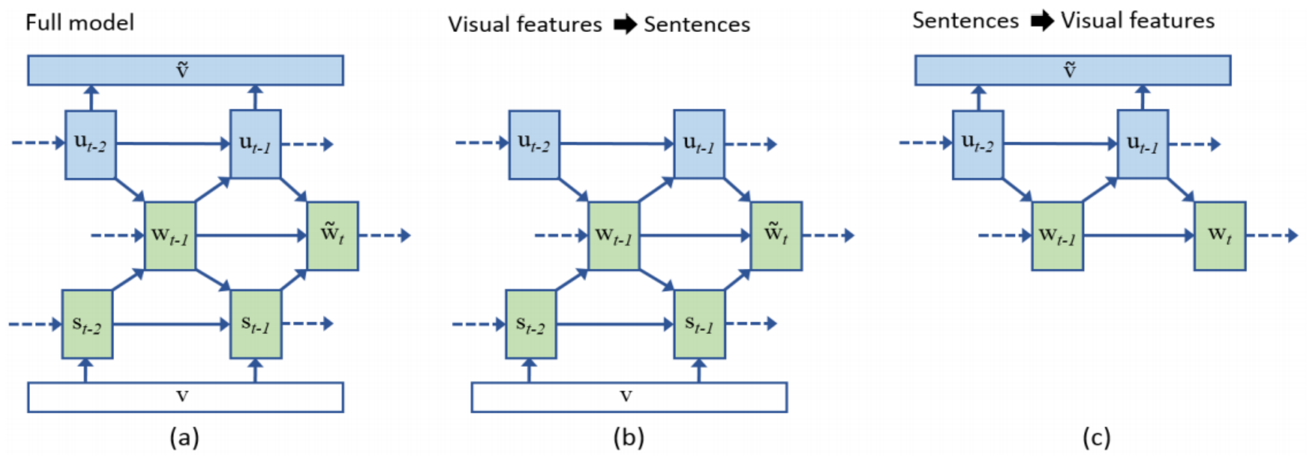
Full model
Visual features ➡ Sentences
Sentences ➡ Visual features
$\tilde{\mathrm{v}}$
$\mathrm{u}_{t\text{-}2}$
$\mathrm{u}_{t\text{-}1}$
$\mathrm{w}_{t\text{-}1}$
$\tilde{\mathrm{w}}_{t}$
w_{t}
$\mathrm{s}_{t\text{-}2}$
$\mathrm{s}_{t\text{-}1}$
v
(a)
(b)
(c)

Figure 2. Illustration of our model. (a) shows the full model used for training. (b) and (c) show the parts of the model needed for generating sentences from visual features and generating visual features from sentences respectively.

together our global objective is to maximize,

$$\begin{aligned} P(w_t, &V|V, W_{t-1}, U_{t-1} \\ &= P(w_t|V, W_{t-1}, U_{t-1}P(V|W_{t-1}, U_{t-1}). \end{aligned} \tag{1}$$

That is, the author want to maximize the likelihood of the word w_t and the observed visual features V given the previous words and their visual interpretation. Note that in previous papers the objective was only to compute $P(w_t|V, W_{t-1}$ and not $P(V|W_{t-1})$.

Following [2], Mikolov *et al.* [3] added an input layer v to the RNN shown by the white box in Fig. 2. This layer may represent a variety of information, such as topic models or parts of speech. Fig. 1 shows an illustrative example of the hidden units **s** and **u**. As can be observed, some visual hidden units **u** exhibit longer temporal stability.

While each word may be predicted independently, this approach is computationally expensive. Instead, the author adopted the idea of word classing and factorized the distribution into a product of two terms:

$$P(w_t|\cdot) = P(c_t|\cdot) \times P(w_t|c_t, \cdot). \tag{2}$$

where $P(w_t|\cdot)$ is the probability of the word, $PP(c_t|\cdot)$ is the probability of the class.

References

[1] A. Gupta, Y. Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012. 1

[2] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014. 2

[3] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 2