Text-Based Image Retrieval Using Deep Learning

1

Udit Singhania
https://orcid.org/0000-0002-4400-2783
Vellore Institute of Technology, India

B. K. Tripathy
Vellore Institute of Technology, India

INTRODUCTION

This chapter is an advanced version of the previous chapter named 'An Insight into Deep Learning Architectures' in the book, 'Encyclopedia of Information Science and Technology, Fourth Edition'. The earlier chapter was submitted by VIT University. It is already an established fact that information retrieval is of two types, image retrieval, and text retrieval. Image retrieval is further divided into a query and content-based approach. The main focus of the chapter will be the same as the earlier version, on Image Retrieval. In the previous edition of this book, the authors attempted to solve information retrieval problems using deep learning architectures. The authors looked on the RBMs, DBNs, and CNNs, how they can be used wisely to solve the issue of image retrieval from queries and text. This chapter will look at some of the advanced architectures (after RBMs and DBNs) used in the text-based approach.

What is information retrieval? Information retrieval means retrieving information from web or data present around us. Information retrieval is of two types image and text; information can be retrieved in either textual form or image form. Google Image search engines are an example of image retrieval. Retrieving documents from text is an example of text retrieval.

Image retrieval is done by deep learning architectures in action. Certainly, CNNs (convolutional neural nets) play an important role, they are the backbone of the image recognition tasks, with the help of NLP (natural language processing techniques) extracting the important features from the text-based query becomes an easy task. Feature learning is an important task while learning for images and text matching, this issue was resolved by using unsupervised approach DBN (Deep Belief Nets) and using new type of architecture was discovered named RBMs (Restricted Boltzmann Machines). After using simple RBMs, there has been a discovery of a new method named Multimodal Learning with Restricted Boltzmann Machines.

What is modality? In this world, information comes through various input channels. Some images have captions and tags, videos have visual and audio signals, sensory perception includes simultaneous inputs from visual, auditory, motor and haptic pathways. Each modality is characterized by separate statistical properties which makes it difficult to disregard the fact that they come from different input channels. Using the combinations of such modalities to jointly represent image and text helps in better models like for example, take an image of a car and words 'A Car', they are assigned a high probability to one conditioned on other. This approach is the base of Multi-modal RBMs.

In the supervised learning side, after CNNs came into existence, there have been dire need of an architecture which can remember past inputs and predict the future input and then came into existence,

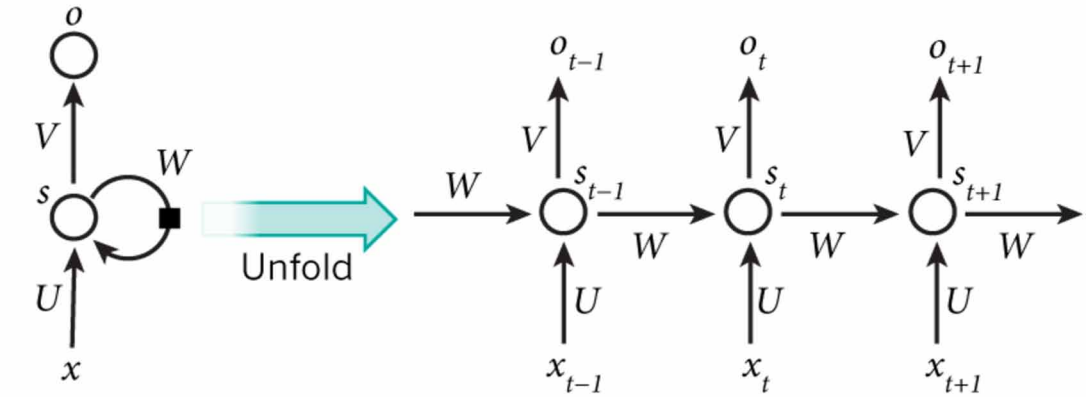
DOI: 10.4018/978-1-7998-3479-3.ch007

new architecture RNNs (Recurrent Neural Nets). RNNs play an important role in textual analysis, time series forecasting, sound and speech recognition.

BACKGROUND

Recurrent neural nets belong to the class of artificial neural networks (ANNs) designed specifically to recognize patterns in the sequences of data, such as text, handwriting, speech, or numerical time-series data which can be gathered from sensors, stock markets, and government agencies. These algorithms have a major focus on time and sequence, looking towards the temporal dimension. They are considered to be one of the most powerful and useful types of a neural network, alongside the attention mechanism and memory networks. RNNs are applicable even to images, which can be decomposed into a series of patches and treated as a sequence. Traditional neural nets could not use the sense of persistence, they failed to perceive the future input because of not able to remember the sequence in which the input was fed, these neural nets have basically two inputs, first they take the input of the previous activation function of the previous input and also consider the present input and they apply activation function of the resultant to get new activation function and predict the output. This can be understood by the figure drawn below.

Figure 1.



Here, RNN is being *unrolled* (or unfolded) into a full network. The formulas that govern the computation happening during a RNN area unit as follows:

- is the input at time step . x_0 is a one-hot vector with respect to 1st word of a sentence.
- is the hidden state at time step . It's the "memory" of the network. is calculated supported the previous hidden state and also the input at this step:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (1)$$

The function "f" usually is a non-linear such as tanh or ReLU (Rectified Linear Unit Function). S_{-1}, the first hidden state, is typically initialized to zero vector.