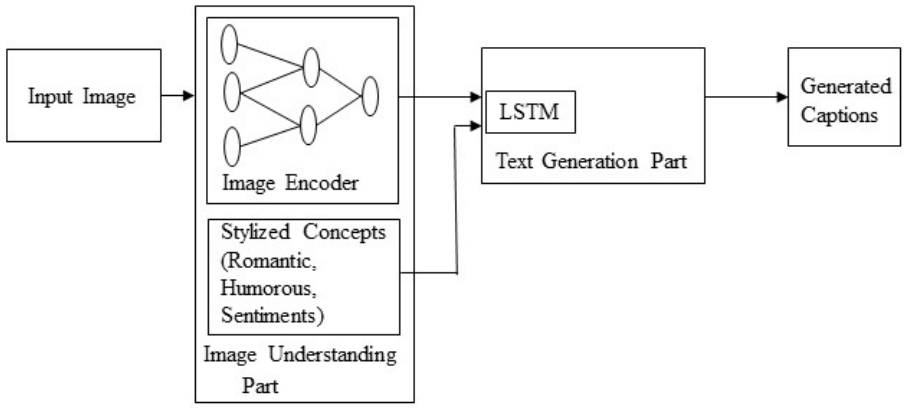

Fig. 10. A block diagram of image captioning based on different styles.

3.5.4 Stylized Caption. Existing image captioning systems generate captions just based on only the image content that can also be called factual description. They do not consider the stylized part of the text separately from other linguistic patterns. However, the stylized captions can be more expressive and attractive than just only the flat description of an image.

The methods of this category follow the following general steps:

(1) CNN based image encoder is used to obtain the image information.
(2) A separate text corpus is prepared to extract various stylized concepts (For example: romantic, humorous) from training data.
(3) The language generation part can generate stylized and attractive captions using the information of Step 1 and Step 2.

A simple block diagram of stylized image captioning is given in Figure 10.

Such captions have become popular because they are particularly valuable for many real-world applications. For example, everyday people are uploading a lot of photos in different social media. The photos need stylized and attractive descriptions. Gan et al. [39] proposed a novel image captioning system called StyleNet. This method can generate attractive captions adding various styles. The architecture of this method consists of a CNN and a factored LSTM that can separate factual and style factors from the captions. It uses multitask sequence to sequence training [89] for identifying the style factors and then add these factors at run time for generating attractive captions. More interestingly, it uses an external monolingual stylized language corpus for training instead of paired images. However, it uses a new stylized image caption dataset called FlickrStyle10k and can generate captions with different styles.

Existing image captioning methods consider the factual description about the objects, scene, and their interactions of an image in generating image captions. In our day to day conversations, communications, interpersonal relationships, and decision making we use various stylized and non-factual expressions such as emotions, pride, and shame. However, Mathews et al. [97] claimed that automatic image descriptions are missing this non-factual aspects. Therefore, they proposed a method called SentiCap. This method can generate image descriptions with positive or negative sentiments. It introduces a novel switching RNN model that combines two CNN+RNNs running in parallel. In each time step, this switching model generates the probability of switching between two RNNs. One generates captions considering the factual words and other considers the words with