# IMAGE CAPTIONING: A SURVEY OF EXISTING ISSUES ON DATASETS, EVALUATION METRICS AND METHODS

by

Liwan Zhou

A THESIS

Submitted to the Faculty of the Stevens Institute of Technology in partial fulfillment of the requirements for the degree of

MASTER OF SICENCE – ELECTRICAL ENGINEERING

_____

Liwan Zhou, Candidate

ADVISORY COMMITTEE

_____

Hong Man, Advisor Date

_____

Cristina Comaniciu, Reader Date

STEVENS INSTITUTE OF TECHNOLOGY

Castle Point on Hudson

Hoboken, NJ 07030

2019

# IMAGE CAPTIONING: A SURVEY OF EXISTING ISSUES ON DATASETS, EVALUATION METRICS AND METHODS

## ABSTRACT

Image captioning is essentially a vision to language problem that combines computer vision and natural language processing. It can be applied to image retrieval, children's education and assisted living for the visually impaired. With the development of deep learning, the approach that combines Deep Convolutional Neural Network and Recurrent Neural Network has made significant progress in image captioning. Although many methods have achieved very high scores in multiple evaluation metrics, judged by human, the results are far inferior to human comprehension. This survey attempts to find out the reasons that cause this situation from the perspectives of datasets, evaluation metrics and methods. We conclude that to better solve the problem of image captioning, we need better evaluation metrics, larger dataset, and new models on both image analysis part and language generation part. Working on these three directions, with the basis of existing image classification capabilities, the image captioning research should be able to come up with methods that are close to or reaching the level of human labeling.

Author: Liwan Zhou

Advisor: Hong Man

Data: December 6, 2019

Department: Electrical and Computer Engineering

Degree: Master of Science in Electrical Engineering

# ACKNOWLEDGMENTS

I would like to express my gratitude to all those who helped me during the writing of this thesis. I gratefully acknowledge the help of my advisor, Prof. Man, who has offered me valuable suggestions in the academic studies.

I also owe a special debt of gratitude to my friend for her constant encouragement.

I should finally like to express my gratitude to my beloved parents who are always by my side.

# Table of Contents

# List of Tables

# List of Figures

# 1. Introduction

Image captioning essentially is a vision to language problem that combines computer vision and natural language processing. It is uncomplicated for a human to understand an image, but machine can not describe image like human. Image captioning is a popular in the field Artificial Intelligence (AI) now.

With the development of deep learning, a method that combines Deep Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) has made significant progress in image captioning (Chen & Zitnick, 2014).

A previous study(Wu, Shen, Liu, Dick, & van den Hengel, 2015) found that their method surpassed human level in the automatic evaluation metric.

Table 1: Results of Att-CNN+LSTM model and human on the test set of COCO database (Wu et al., 2015)

| COCO-TEST | B-1 | B-2 | B-3 | B-4 | M | R | CIDEr |
|---|---|---|---|---|---|---|---|
| **5-Refs** | | | | | | | |
| Att-CNN+LSTM | **0.73** | **0.56** | **0.41** | **0.31** | **0.25** | **0.53** | **0.92** |
| Human | 0.66 | 0.47 | 0.32 | 0.22 | **0.25** | 0.48 | 0.85 |
| **40-Refs** | | | | | | | |
| Att-CNN+LSTM | **0.89** | **0.80** | **0.69** | **0.58** | 0.33 | **0.67** | **0.93** |
| Human | 0.88 | 0.74 | 0.63 | 0.47 | **0.34** | 0.63 | 0.91 |

5-Refs and 40-Refs indicate that there are two data sets in the test set, one data set has 5 reference labels per image, which is the correct sentence input by humans, and one data set has 40 reference labels.

B-N (N = 1, 2, 3, 4), M, R and CIDEr represent 4 different automatic evaluation standards for the algorithm, which will be described in chapter 4. In simple terms, the higher the score, the better the model is. Participated in the comparison are **Att-CNN+LSTM** model(Wu et al., 2015) and human level. The table shows that 13 of the thesis methods have exceeded human scores of the 14 scores. However, it is obvious that machine can not be more precise than human in describing images with current level, and we will discuss the results.

Although the highest level of image caption algorithm has not yet surpassed human performance, the breakthrough of it is beyond doubt. There are some methods giving some good caption that human can admit.

A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surfboard.

Figure 1: Some good example of image captioning

If the problem of image captioning can be solved well, it can be applied to image retrieval, children's education and assisted living for the visually impaired. From an academic perspective, the current research on image caption issues has prompted the two fields of artificial intelligence, computer vision and natural language processing to combine well. This cross-sub-field combination can lead to more amazing method.

# 2. Literature review

The image caption problem has been studied for many years. The datasets have changed a lot. From the Flickr8K and 30K that have been popular for several years to MS COCO dataset (Chen et al., 2015), the dataset becomes bigger, besides, COCO dataset has sub datasets that each image is with man-made generated sentences which can make evaulation easier. Chen and Zitnick (2015) also think that fine-tuning strategy is particularly helpful for large dataset instead of small dataset.

Evaluations performed over machine generated description of image can be divided into Automatic Evaluation and Human Evaluation. There are five different metrics using for automatic evaluation: Perlexity (Chen & Zitnick, 2014), BLEU (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015) and ROUGE (Lin, 2004). For automatic evaluation, when comparing the generating sentences with human reference sentences, these metrics give higher score to machine rather than human in some methods (Wu et al., 2015), which makes these automatic evaluaiton metrics less reliable.

Mao, Xu, Yang, Wang and Yuille (2014) proposed multimodal Recurrent Neural Network model that connects a CNN to an RNN to learn the mapping from image to sentences. A improved model that changed CNN to LSTM (Vinyals, Toshev, Bengio, & Erhan, 2015) were appeared. Then a new idea called o Explicit High Level Concepts (Wu et al., 2015) come out and achieved a significant improvement.

The main contribution of this paper is looking through the development progress of datasets, evaluation metrics and image caption methods and finding out what the breakthrough we can make in the future.

# 3: Datasets

There has been no breakthrough in small sample learning, so deep learning is inseparable from big data. In the process of image caption research, researchers' preference for the database is also changing.

## 3.1 Microsoft COCO Dataset

The generation of the Microsoft COCO Caption dataset is based on the work of the Microsoft Common Objects in COntext (COCO) dataset.

The COCO dataset contains two sub datasets: one is MS COCO c5. The training set, validation set and test set images it contains are consistent with the original MS COCO database, except that each image is accompanied by 5 manually generated caption sentences. The other is MS COCO c40. It contains only 5000 images, and these images were randomly selected from the test set of the MS COCO dataset. Different from c5, each of its images has 40 man-made caption sentences.

The other job the creators have done is to set up an evaluation server to achieve the most popular evaluation standards (BLEU, METEOR, ROUGE, and CIDEr). Upload your sentences generated by your method to the server for the test set images, and the server will automatically give scores for various evaluation.

## 3.2 Flickr8K and 30K

The source of the image data is Yahoo's photo album website Flickr and the number of images in the dataset is 8,000 and 31,783 respectively.

Most of the images in these two databases show human participation in an event. The corresponding manual label for each image is 5 sentences. The syntax of the annotations for

these two databases is similar. The database is also divided into multiple blocks based on the standard training and validation test sets.

Compared with the COCO dataset, the obvious disadvantage of the Flickr8K and Flickr30K datasets is the insufficient data volume. The shortcomings of data volume do make Flickr datasets gradually lose competitiveness. In the past, almost all papers have used this dataset. Some advanced papers now only display it in supplementary documents or even not use it.

### 3.3 PASCAL 1K

this dataset is a subset of the PASCAL VOC challenge image dataset. For its 20 classifications, 50 images are randomly selected, for a total of 1,000 images. Then the same applies to Amazon's Turkish robot service, which manually labels 5 description sentences for each image. Generally speaking, this dataset is only used for testing.

# 4. Evaluation metrics

In introduction, table 1 indicated that the model has scored more than human score in the multiple evaluation metric, however the true level of machine can not exceed human. The reason for this situation is automatic evaluation metric.

## 4.1 Human evaluation and automatic evaluation metric

The most reliable and authoritative evaluation of the quality of the label sentence generated by the algorithm is human. At present, all kinds of automatic evaluation metrics try to make their calculation results relevant to human judgment results.

Table 2: Spearman's correlation co-efficient of automatic evaluation measures against human judgements (Elliott & Keller, 2014)

| | Flickr 8K co-efficient n=17,466 | E&K(2013) co-efficient n=2,040 |
|---|---|---|
| METEOR | 0.524 | 0.233 |
| ROUGE SU-4 | 0.435 | 0.188 |
| Smoothed BLEU | 0.429 | 0.177 |
| Unigram BLEU | 0.345 | 0.097 |
| TER | -0.179 | -0.044 |

The correlation coefficient with human judgment is not correlated between 0.0–0.1, 0.11–0.4 is weakly correlated, 0.41–0.7 is moderately correlated, 0.71–0.90 is strongly correlated, and if it is 0.91-1.0, it is perfect. So the conclusion is to first recommend METEOR, or use ROUGE SU-4 and Smoothed BLEU(Elliott & Keller, 2014).

The experiment used Microsoft's COCO dataset. Among the three evaluation criteria scores, the score of the Google NIC model and the human score are inconsistent.

Table 3: Score on the MSCOCO development set

| Metric | BLEU-4 | METEROR | CIDER |
|---|---|---|---|
| NIC | **27.7** | **23.7** | **85.5** |
| Random | 4.6 | 9.0 | 5.1 |
| Nearest Neighbor | 9.9 | 15.7 | 36.5 |
| Human | 21.7 | 25.2 | 85.4 |

In figure 2, The x-coordinate is the BLEU score and the y-coordinate is the cumulative distribution, which is the percentage of the output description statement set is greater than the current x's score. And among them:

- Flickr-8k: NIC represents the scoring curve of the results of running on the Flick8k test set using the NIC model;
- Pascal: NIC represents the scoring curve of the results of running on the Pascal test set using the NIC model;
- COCO-1k: NIC represents the scoring curve of the results of running on the COCO-1k test set using the NIC model;
- Flickr-8k: ref represents the score curve of the results of another paper, which is used as a benchmark;

- Flickr-8k: GT represents the result of artificially graded evaluation of the artificially labeled sentences of Flickr-8k images.

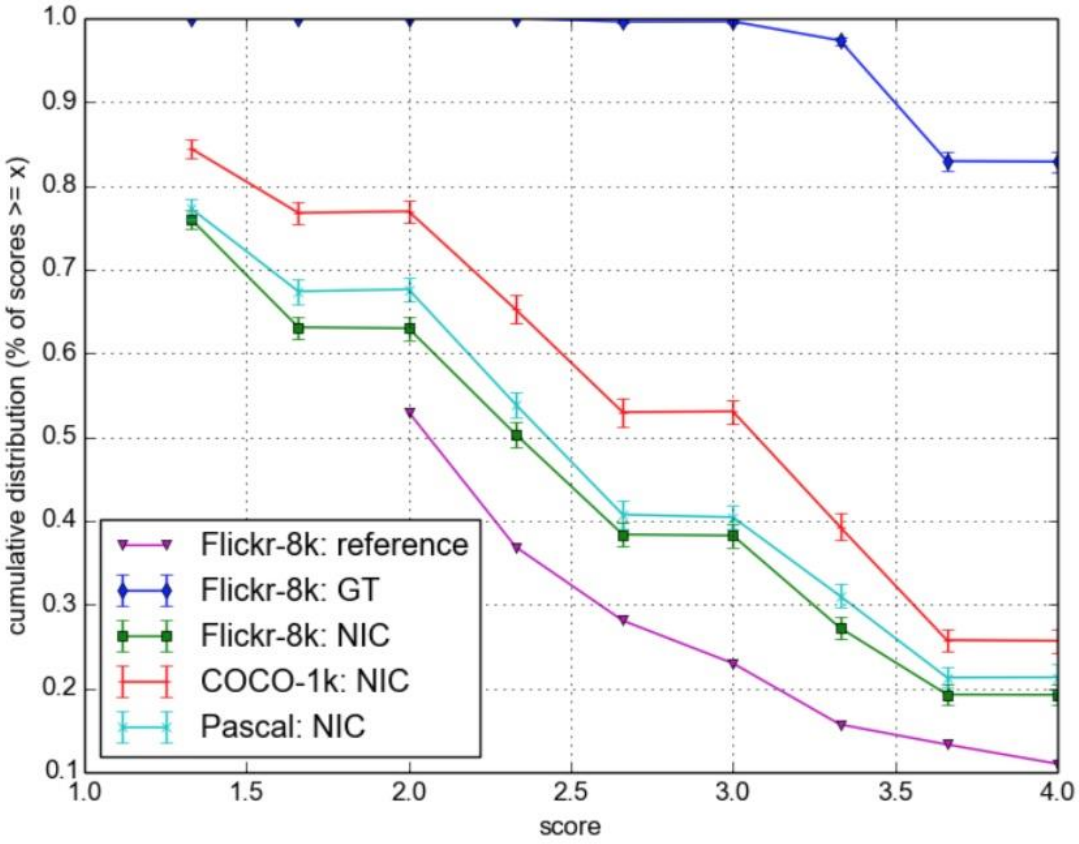


Figure 2: Comparison of evaluation of multiple methods (Vinyals et al., 2015).

Although the automatic metric BLEU-4 believes that the score of the NIC model exceeds that of humans, the NIC model is far inferior to human comprehension judged by human evaluation.

## 4.2 Perplexity

This evaluation metric is first used in natural language processing. In a study(Mao, Xu, Yang, Wang, & Yuille, 2014), which was first proposed to combine RNN and CNN model, perplexity was defined as: $\log_2 PPL\left(w_{1:L} \mid I\right) = -\frac{1}{L}\sum_{n=1}^{L}\log_2 P\left(w_n \mid w_{1:n-1}, I\right)$, where $\mathrm{L}$ is the length of the sentence, and $PPL\left(w_{1:L} \mid I\right)$ is the perplexity of the sentence $w_{1:L}$ given according to the image $I$ . $P\left(w_n \mid w_{1:n-1}, I\right)$ is the probability of generating the next word $w_n$ based on the image $I$ and the previous word sequence $w_{1:n-1}$ .

From this define formula, the perplexity value will increase when the model's certainty for the next generated word decreases. In other word, if the complexity of language model is low enough, a machine can write fluent and logical sentences.

## 4.3 BLEU

BLEU (Bilingual Evaluation Understudy) is widely used in the evaluation of image caption results, but it was originally designed for machine translation problems. It is used to analyze the correlation of n-tuples between the translation sentence to be evaluated and the reference translation sentence. Its core idea is: the closer a machine translation is to a professional human translation, the better it is (Papineni et al., 2002).

For the image $I_i$, the model will generate corresponding annotation statement $c_i$, and the automatic evaluation metric can evaluate the quality of the evaluation standard statement $c_i$,according to a set $S_i = \{s_{i1}, \ldots, s_{im}\} \in S$ of reference annotation statement. Label statements are expressed in n-grams. An n-tuple $w_k \in \Omega$ is a sequence of one or more sequential words.

Now we generally only explore n-tuples from 1 word to 4 words. The number of occurrences of an n-tuple $w_k$ in a statement $s_{ij}$ is denoted as $h_k\left(s_{ij}\right)$, and the number of occurrences of an n-tuple $w_k$ in a statement $c_i \in C$ to be evaluated is denoted as $h_k\left(c_i\right)$.

The calculation formula of BLEU is as follows:

$$CP_n(C,S)=\frac{\sum_i\sum_k \min\left(h_k\left(c_i\right),\max_{j\in m} h_k\left(s_{ij}\right)\right)}{\sum_i\sum_k h_k\left(c_i\right)}$$

The global n-tuple precision is first calculated: where k refers to the set number of possible n-tuples of length n.

However, in practice, it is not ideal to use a single word for comparison, so BLEU uses n-tuples to calculate, and the value of n is up to 4. The number of 1-tuples is not enough to evaluate translation. Longer tuple scores correspond to fluency.

Then a conciseness penalty is introduced, because BLEU tends to shorter sentences, the accuracy score will be high. To solve this problem, multiplication by a concise penalty is used to prevent very short sentences from getting high scores.

The penalty value for conciseness is then calculated: where the formula is the total length of the statements to be evaluated, and the total length of a globally valid reference sentence. If there are multiple reference statements for a statement to evaluate, choose the one that has the least penalty for brevity.

$$b(C,S)=\begin{cases}1 & \text{if} \quad l_C > l_S \\ e^{1-l_S/l_C} & \text{if} \quad l_C \le l_S\end{cases}$$

Let $l_S$ be the total length of the reference sentence, and $l_C$ be the total length of the sentence to be evaluated. If $l_C$ is less than or equal to $l_S$, the penalty takes effect and $e^{1-l_S/l_C}$ is calculated. Conversely, the simplicity penalty is 1. If there are multiple reference sentences, the length of the reference sentence that is closest to the length of the sentence to be evaluated is selected.

## 4.4 ROUGE

ROUGE is a set of automatic evaluation measure designed to evaluate text summarization algorithms. There are three evaluation measures, namely ROUGE-N, ROUGE-L and ROUGE-S.

The first one is ROUGE-N. Based on the sentence to be evaluated; it calculates a simple n-tuple recall for all reference summaries:

$$\mathrm{ROUGE}_n\left(c_i, S_i\right) = \frac{\sum_j \sum_k \min\left(h_k\left(c_i\right), h_k\left(s_{ij}\right)\right)}{\sum_i \sum_k h_k\left(s_{ij}\right)}$$

ROUGE-L is a measurement method based on longest common subsequence (LCS). The so-called LCS is a set of words that appear in two sentences at the same time, and the order in which the words appear is the same. Unlike n-tuples, there may be words that can create LCS between words. Let the length of the LCS between the two sentences compared be: $l\left(c_i, s_{ij}\right)$. By calculating F1 score, ROUGE-L can get:

$$R_l = \max_j \frac{l\left(c_i, s_{ij}\right)}{\left|s_{ij}\right|},\quad P_l = \max_j \frac{l\left(c_i, s_{ij}\right)}{\left|c_i\right|},\quad \mathrm{ROUGE}_L\left(c_i, S_i\right) = \frac{\left(1+\beta^2\right) R_l P_l}{R_l + \beta^2 P_l}$$

$R_l$ is the recall, $P_l$ is the precision, $\beta$ is generally equal to 1.2, and n-tuples is irrelevant in this calculation.

ROUGE-S is the last criterion, it uses skip bigram. Jumping tuples are ordered pairs of words in a sentence, and similarity to LCS, words can be skipped between pairs of words. For example, a sentence with 4 words may have 6 types of skipping tuples in permutations and combinations. Calculate F1-score[6] again using precision and recall, and record the number of skipping tuples in the sentence $s_{ij}$ as $f_k\left(s_{ij}\right)$, then the calculation formula is as follows:

$$R_s = \max_j \frac{\sum_k \min\left(f_k\left(c_i\right), f_k\left(s_{ij}\right)\right)}{\sum_k f_k\left(s_{ij}\right)},\ P_s = \max_j \frac{\sum_k \min\left(f_k\left(c_i\right), f_k\left(s_{ij}\right)\right)}{\sum_k f_k\left(c_i\right)},$$

$$\mathrm{GOUGE}_S\left(c_i, S_i\right) = \frac{\left(1+\beta^2\right) R_s P_s}{R_s + \beta^2 P_s}$$

## 4.5 METEOR

METEOR is a metric used to evaluate machine translation output. This method is based on the precision of a single tuple and the harmonic mean [7] of recall. The weight of the recall is slightly higher than the precision. This metric has some features that other metrics do not have. It is designed to solve some problems of BLEU. It is highly relevant to human judgment, and different from BLEU, it is highly related to human judgment not only at the word set, but also at the sentence and segment level. At the ensemble level, its correlation is 0.964 and BLEU is 0.817. At the sentence level, its relevance is up to 0.403.

METEOR's calculation formula:

$$Pen = \gamma\left(\frac{ch}{m}\right)^{\theta},\ F_{\text{mean}} = \frac{P_m R_m}{\alpha P_m + (1-\alpha)R_m},\ P_m = \frac{|m|}{\sum_k h_k(c_i)},\ R_m = \frac{|m|}{\sum_k h_k(s_{ij})},$$

$$METEOR = (1-Pen)F_{mean}$$

where $m$ is the set of planar alignments, $ch$ is the number of chunks, $P_m$ is the precision, and $R_m$ is the recall rate.

### 4.6 CIDEr

CIDEr is specifically designed for image captioning. It measures the consistency of image captioning by calculating the Term Frequency Inverse Document Frequency (TF-IDF) weight for each n-tuple.

The number of times an n-tuple $w_k$ appears in the reference sentence $s_{ij}$ is recorded as $h_k(s_{ij})$, and if it appears in the sentence to be evaluated, it is recorded as $h_k(c_i)$. CIDEr calculates TF-IDF weights $g_k(s_{ij})$ for each n-tuple $w_k$:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega}} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right)$$

where Ω is the and I is the set of all images in dataset. The first part of the formula calculate the TF of every n-tuple $w_k$, the second part use IDF to calculate the rarity of $w_k$. That is, if some n-tuples appear in the reference label of image captioning frequently, TF will give these n-tuples higher weight while IDF will lower the weight of n-tuples that appear in all description statements frequently.

# 5. Image caption methods

In the development of image caption methods, the core idea that combining the RNN and CNN has not changed.

## 5.1 m-RNN model

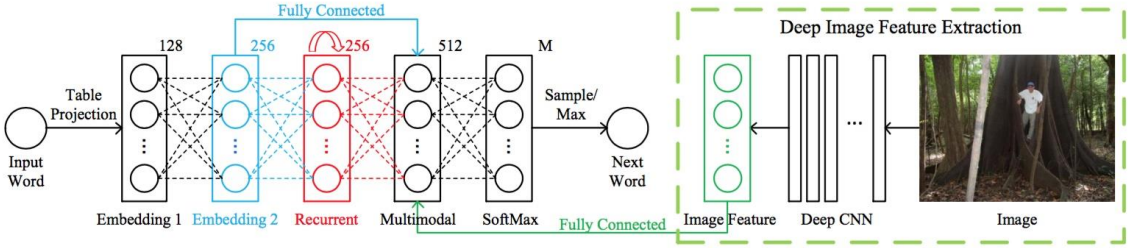


Figure 3: The structure of m-RNN model (Mao et al., 2014)

Its structural features can be summarized as follows:

The input to the model is the image and the annotation statement corresponding to the image (for example, in the figure above, the statement might be a man at a giant tree in the jungle). The output is the distribution of possibilities for the next word;

Each time frame of the model has six layers: the input layer, two word embedding layer, the recurrent layer, the multimodel layer and the SoftMax layer.

The input words were originally encoded in one-hot, but after two word embedding layer, they were finally transformed into dense words. The word expression layer is randomly initialized and learned by itself during training. The activation data output from the second embedding layer is directly entered into the multimodal layer as input;

The dimension of the recurrent layer is 256, in which the transformation and calculation of the word expression vector $w(t)$ at time t and the activation data $r(t-1)$ at time t-1 are carried

out. The specific calculation formula is: $r(t) = f_2\left(U_r \cdot r(t-1) + w(t)\right)$ , where the functio $f_2\left(\cdot\right)$ is ReLU, and $U_r$ is the transformation to map $r(t-1)$ to the same vector space as $w(t)$;

The 512-dimensional multimodal layer connects the language part and the image part of the model. The image part essentially uses the deep convolutional neural network to extract the image features. The activation data of the seventh layer of AlexNet is used as the feature data input to the multimodal layer so the image feature vector I is obtained. The language part contains the word embedding layer and the recurrent layer;

The calculation in the multimodal layer is: $m(t) = g_2\left(V_w \cdot w(t) + V_r \cdot r(t) + I\right)$, where m represents the feature vector of the multimodal layer, I represents the feature vector of the input part of the image, and the interpretation of $w(t)$ and $r(t)$ is the same as above. As for the $V_w$ and the $V_r$, they are matrix transformations. In this formula: at each time t, the image feature $I$ is entered into the calculation as input. Finally, the $g_2\left(\cdot\right)$ function is a tanh function with parameters: $g_2(x) = 1.7159 \cdot \tanh\left(\frac{2}{3}x\right)$. The cost function is designed based on the Perplexity when the model is trained:

$$C = \frac{1}{N}\sum_{i=1}^{N} L \cdot \log_2 PPL\left(w_{1:L}^{(i)} \mid I^{(i)}\right) + \| \theta\|_2^2$$

where N is the number of words in the training set, $\theta$ is the parameter of the model and $\| \theta\|_2^2$ is a regularization part. And L is the length of the sequence of words. The goal of training is to minimize the cost function. We can use stochastic gradient descent to learn the parameters.

**Statement generation of the model**: the model starts with a special start symbol "##START##" or any reference word and then the model calculates the probability distribution of

the next word $P\left(w\middle|w_{1:n-1}\middle|I\right)$. Then taking the word with the highest probability as the selected word, using the word as the input to predict the next word, and repeat until the end symbol ##END## is generated.

**Dataset and evaluation**: Because the model was proposed early, the datasets used to this model are Flickr8K, 30K, and IAPR tc-12. And there are also few the automatic evaluation metrics used, with only Perplexity and BLUE1-3.

This is the first model combining RNN and CNN. The following models are basically optimized on the basis of this model.

## 5.2 NIC model

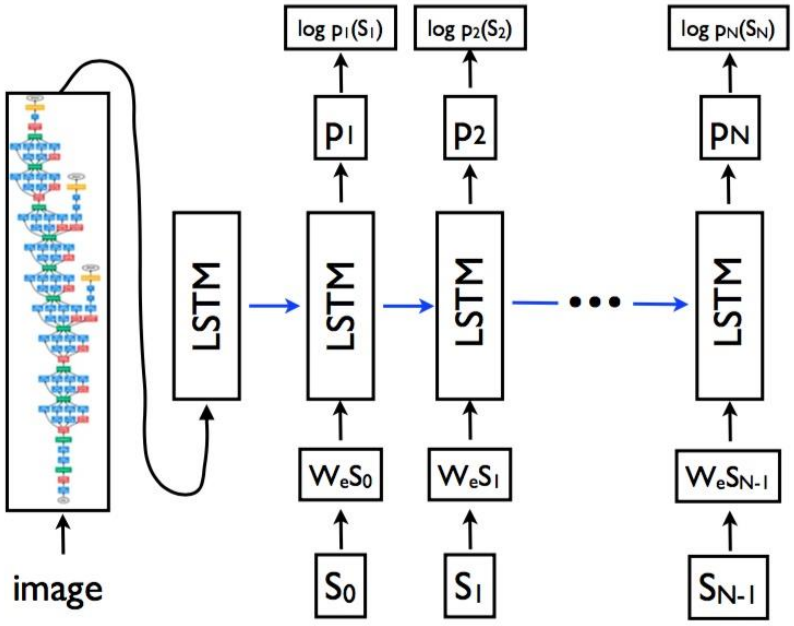


Figure 4: The structure of NIC model (Vinyals et al., 2015)

Compared with the m-RNN model, the NIC model has the following differences:

- Discard RNN and use LSTM;
- The CNN part uses convolutional neural network better than AlexNet;
- The image feature data extracted by CNN is input only once at the beginning.

The part about image feature is almost the same: the image passes through the convolutional neural network and eventually becomes the feature vectors. The only difference is that the CNN is different, and the image feature is only input to LSTM at the beginning.

The difference in the language part is that NIC model uses LSTM instead of RNN.

The process shown in the above model can be summarized as:

$$x_{-1} = CNN(I)$$
$$x_t = W_e S_t, \quad t \in \{0 \ldots N-1\}$$
$$p_{t+1} = LSTM\left(x_t\right), \quad t \in \{0 \ldots N-1\}$$

And the cost function is: $L(I,S) = -\sum_{t=1}^{N} \log p_t\left(s_t\right)$

After these improvements, we can figure out the NIC model is more advanced than the m-RNN model from the results.

Table 4: Comparison of BLEU-1 scores of different algorithms on different datasets

| Approach | PASCAL(xfer) | Flickr 30k | Flickr 8k | SBU |
|---|---|---|---|---|
| m-RNN | - | 55 | 58 | - |
| NIC | 59 | 66 | 63 | 28 |
| Human | 69 | 68 | 70 | - |

**In summary**: compared with the m-RNN model, the NIC model has the following three improvements:

- First, in the language model section, RNN is replaced by LSTM, which is proved to be more effective in NLP.
- Secondly, a better convolutional neural network model is used in the image part to extract the image feature data.
- Finally, the input mode of image feature data was changed from every time point of m-RNN to only one time at the initial time.

## 5.3 Att+CNN+LSTM model

Once familiar with the CNN + RNN model, we know that the results will come out when the image features pass through the model, however, one important step is missing in the process. A previous paper indicates that this approach does not explicitly represent high-level semantic concepts, but rather seeks to progress directly from image features to text. This paper proposes a method of incorporating high-level concepts into the successful CNN + RNN approach(Wu et al., 2015).

The main improvement is in the vision understanding part. First, pre-train a single-label CNN, then transfer the parameters of this CNN to the multi-label CNN below, and fine-tune the multi-label CNN.

The image is input to multi-label CNN, and the output is a vector $V_{att}(I)$ with a high-level concept and corresponding probability, which is taken as the input of the language part LSTM.

The most valuable part of this paper is how to transfer image to attribute in its image analysis part, which is its core innovation point.

**Construction of the attribute vocabulary**: semantic attributes are extracted from the tag statement of the training set and can be any part of the sentence: noun, verb or adjective. The attribute vocabulary is constructed using c of the most commonly used words. In construction, plurals and tenses are indistinguishable, this effectively reduces the number of vocabularies, resulting in an attribute vocabulary of 256 words.

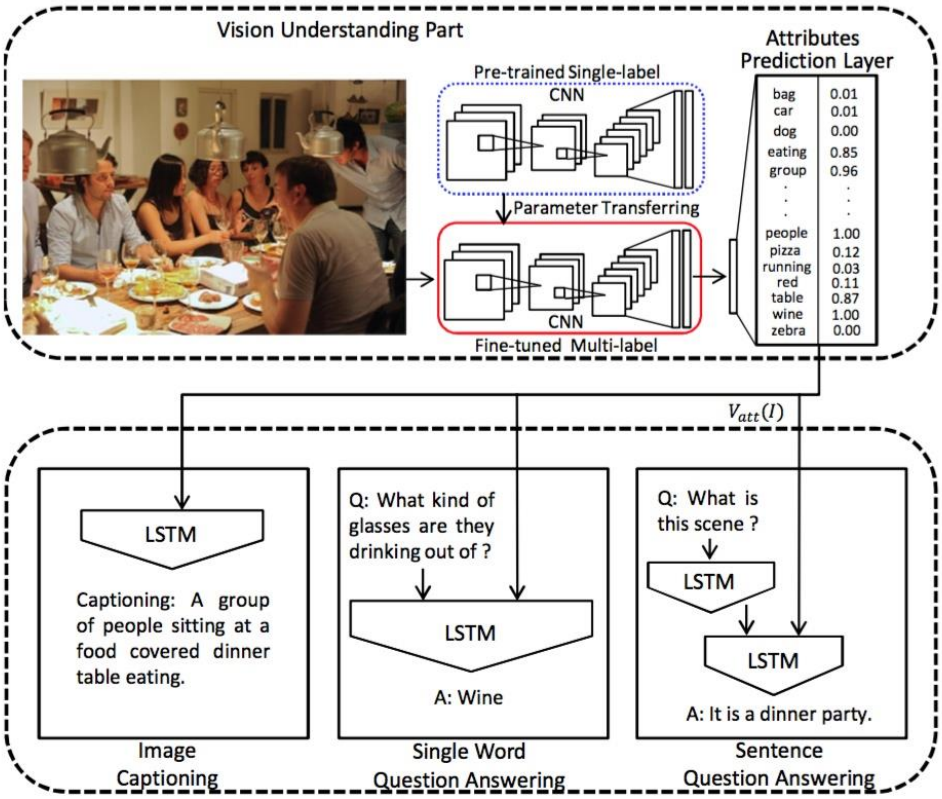


Figure 5: The structure of Att-CNN-LSTM model (Wu et al., 2015)

**Implementation of attribute predictor**: with an attribute vocabulary, we can give a picture and get multiple corresponding attribute words in the vocabulary. This can be considered as a multi-label classification problem. The specific implementation process is shown in the figure 6.

First, take a pre-trained VGGNet(Simonyan & Zisserman, 2014) model with ImageNet as the initial model. Then, the VGGNet can be fine-tuned with multi-label datasets like MS COCO.

Suppose there are N training samples, and $y_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]$ is the label vector corresponding to the *i*th image. If $y_{ij}$=1, it means that the label is present in the image; otherwise, it is not. $p_i = [p_{i1}, p_{i2}, \ldots, p_{ic}]$ is the corresponding prediction probability vector, then the loss function is: $J = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\log\left(1+\exp\left(-y_{ij}p_{ij}\right)\right)$.

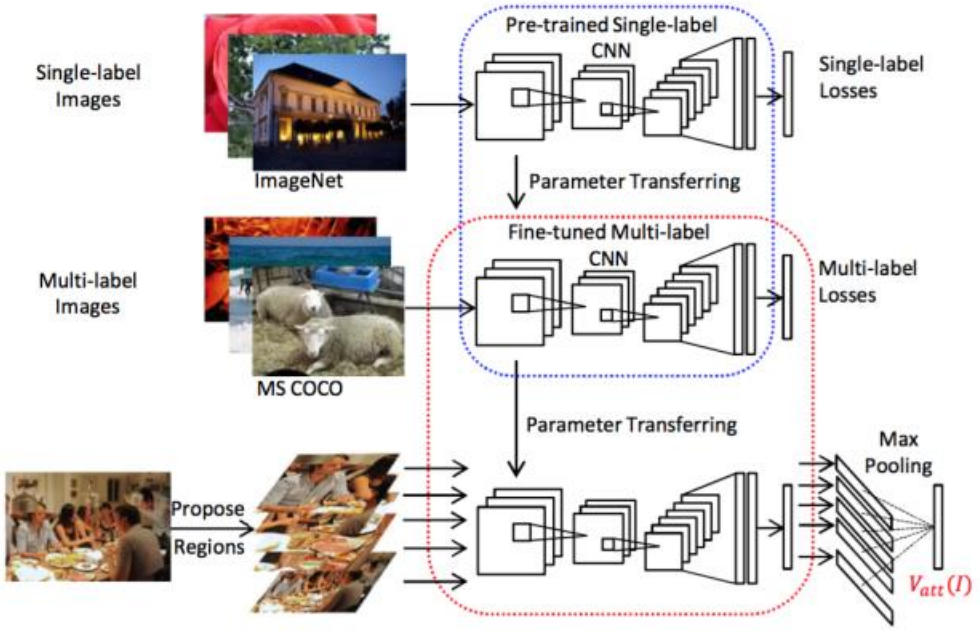


Figure 6: The structure of vision standing part (Wu et al., 2015)

The experimental result(Wu et al., 2015) shows this method is improved significantly.

And from the table 1 in introduction, we have already known that the fact that the score of automatic evaluation criterion is higher than that of human does not mean that the actual annotated statement is better than the reference.

Table 5: BLEU-1,2,3,4, METEOR, CIDEr and PPL metrics compared with other methods and method in this paper on MS COCO dataset (Wu et al., 2015)

| Method | B-1 | B-2 | B-3 | B-4 | METEOR | CIDEr | Perplexity |
|---|---|---|---|---|---|---|---|
| NeuralTalk | 0.63 | 0.45 | 0.32 | 0.23 | 0.20 | 0.66 | - |
| NIC | - | - | - | 0.28 | 0.24 | 0.86 | - |
| LRCN | 0.67 | 0.49 | 0.35 | 0.25 | - | - | - |
| VNet+LSTM | 0.61 | 0.42 | 0.28 | 0.19 | 0.19 | 0.56 | 13.58 |
| VNet-PCA+LSTM | 0.62 | 0.43 | 0.29 | 0.19 | 0.20 | 0.60 | 13.02 |
| GNet+LSTM | 0.60 | 0.40 | 0.26 | 0.17 | 0.19 | 0.55 | 14.01 |
| VNet+ft+LSTM | 0.68 | 0.5 | 0.37 | 0.25 | 0.22 | 0.73 | 13.29 |
| Att-GT+LSTM | 0.80 | 0.64 | 0.50 | 0.40 | 0.28 | 1.07 | 9.60 |
| Att-SVM+LSTM | 0.69 | 0.52 | 0.38 | 0.28 | 0.23 | 0.82 | 12.62 |
| Att-CNN+LSTM | 0.74 | 0.56 | 0.42 | 0.31 | 0.26 | 0.94 | 10.49 |

From the evolution of these three models, we can see a clear process: when a new model is put forward, this model will continue to be improved partly, exploring unclear parts of the new model and the big boost will happen in these blurry areas.

# 6. Conclusion

We presented a survey about the popular dataset, automatic evaluation metrics and methods of the image captioning. The results show that the performance of existing method and the limit of the datasets and untrustworthiness of automatic evaluation metrics. At the end, we indicate future research direction that can better solve the image caption problems.

COCO dataset is best dataset in the field of image caption, because it is a big dataset with annotated sentences corresponding to images. As mentioned in the dataset chapter, big datasets are important for deep learning-based algorithms. Therefore, the improvement of datasets is more images and tagged sentences.

Our survey shows that one main bottleneck obstructing this research progress is the lack of effective image description evaluation metrics. effective automatic evaluation metrics are those that are more relevant to human evaluation. One method to improve these indicators is to increase the number of reference sentences.

The main challenge comes from the image caption method, although the explicit high level concepts bring a significant improvement, other parts can be further improved. For instance, new models or ideas appear in the image analysis part, the language generation part, or the connection between the two parts.

# References


Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR), 52*(6), 115.

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Paper presented at the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/abs/2015arXiv150400325C

Chen, X., & Zitnick, C. L. (2014). Learning a Recurrent Visual Representation for Image Caption Generation. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/abs/2014arXiv1411.5654C

Elliott, D., & Keller, F. (2014). *Comparing automatic evaluation measures for image description.* Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Karpathy, A., & Fei-Fei, L. (2015). *Deep visual-semantic alignments for generating image descriptions.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries.* Paper presented at the Text summarization branches out.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). *Cider: Consensus-based image description evaluation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Wu, Q., Shen, C., Liu, L., Dick, A., & van den Hengel, A. (2015). What value do explicit high level concepts have in vision to language problems? *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/abs/2015arXiv150601144W