Course: Introduction to Deep Learning

MIDTERM PROJECT

Image Captioning
CNN-LSTM based and Attention Mechanism Applications

Instructor: Prof. Le Anh Cuong

Group:

- 520K0127 – Do Pham Quang Hung
- 520K0343 – Le Phuoc Thinh
- 5190K0078 – Chibuike Timothy Benedict

Outline

I. Introduction

II. The earlier image caption methods

III. The recent deep learning methods

IV. Comparison of state-of-the-art methods

V. Experiment conduction result

VI. References

Outline

I. Introduction

II. The earlier image caption methods

III. The recent deep learning methods

IV. Comparison of state-of-the-art methods

V. Experiment conduction result

VI. References

I. Introduction

Image captioning means that given an image, the machine perceives its content and generates descriptions automatically.

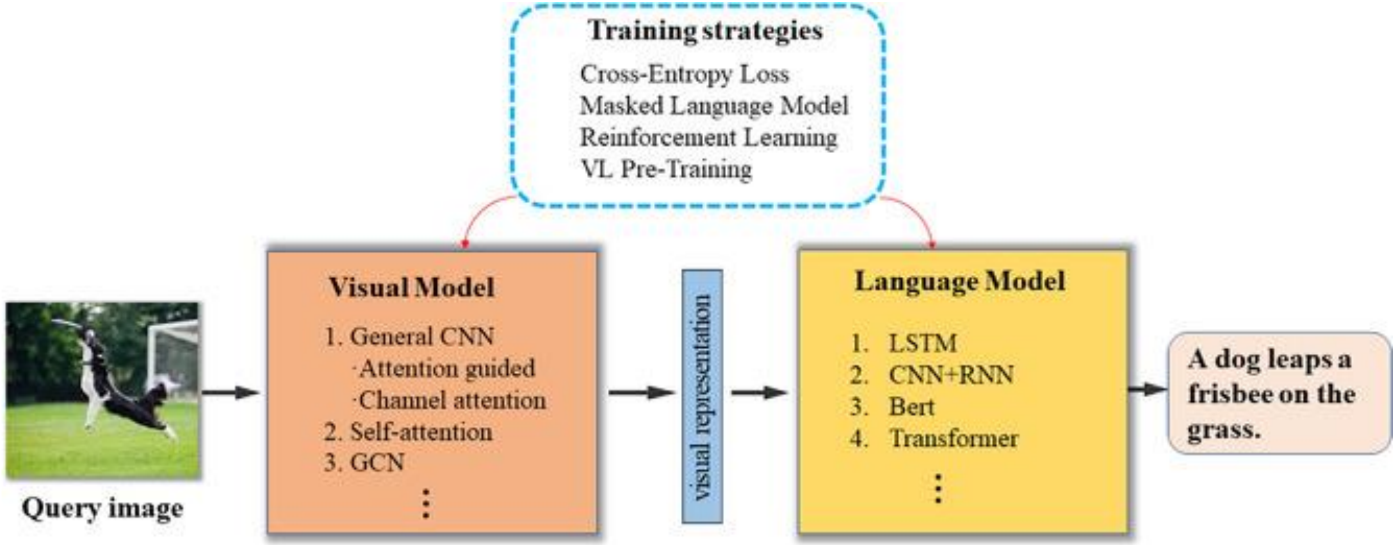
- In the early days of the development of computer vision, researchers tried to use computers to simulate the human visual system and let the computer tell people what it saw. After that, researchers put forward higher requirements:
  - *let the computer recognize the objects in the image, determine the target attributes, and even determine the relationship between the recognized entities in the form of natural language to describe the image.*

So far, there have been many related methods of captioning, and still a continuous improvement.

I. Introduction

The figure below gives an overview of automatic image captioning tasks and a simple example of the most relevant approaches. The purpose of these studies is to find an effective pipeline to process the query image, represent its content, and transform it into a sequence of words by generating connections between visual and textual elements while maintaining the fluency of the language

- In its standard configuration, **image captioning is an image-to-sequence issue.** These images are coded into one or more feature vectors in the visual coding step, and the input is prepared for the second decoder generation step, called a language model. A sequence of words or sub-words decoded from a given vocabulary through a decoder.



Outline

I. Introduction

II. The earlier image caption methods

III. The recent deep learning methods

IV. Comparison of state-of-the-art methods

V. Experiment conduction result

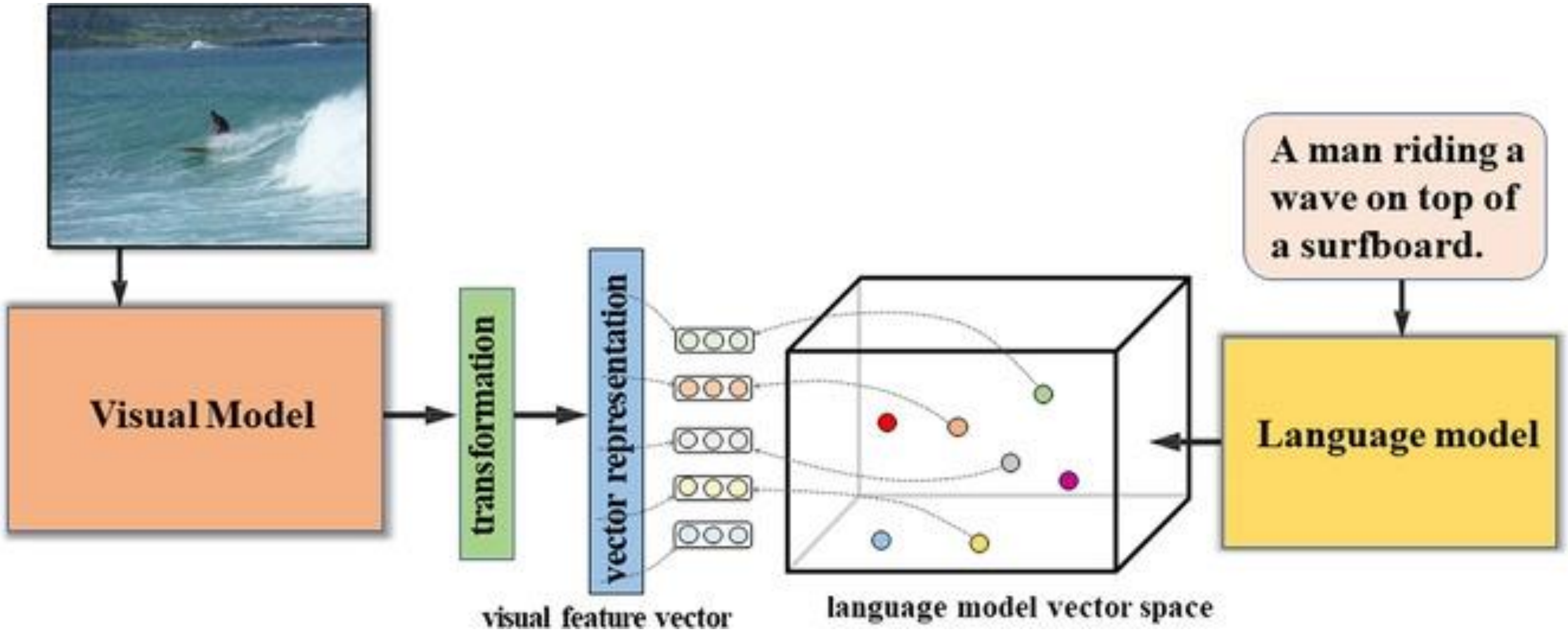
VI. References

II. The earlier image caption methods

This section introduces the two earlier methods: **the search-based method** and **the language template-based method**. Both of them first extract image visual features, such as objects, actions, relationships, scenes, and so forth, across visual models, and then transform them into vector representations.

II. The earlier image caption methods

The difference is their subsequent steps. The **search-based method** is to search for similar images through key information of the image and use the corresponding description as the final description result. While the **language template-based method** maps the image description and the vectorized image content to the same metric space through the language template and selects the result with the highest similarity as its final description.



II. The earlier image caption methods

2.1. Search-based Method

- **Search-based image captioning** approaches usually *construct an image and the corresponding caption into an "image-caption" dataset firstly*. When performing an image caption task, compare the query image with the image in the training set to search for similar images in the training set. Then, the captions of similar images in the dataset are marked as candidate captions set, which are usually sentences or phrases. Finally, the final image description is determined by re-ranking all candidate descriptions.

II. The earlier image caption methods

2.1. Search-based Method

Representative works

Ordonez, V., Kulkarni G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 24, 1143–1151 (2011)

Ordonez et al. grabbed a large number of images from the Internet and manually labeled the title and caption of images. When performing the description task, they **calculated the global similarity (scene) between the image to be described and the image in the network image library**. The most similar images are found, and their corresponding captions are used as the final result.

II. The earlier image caption methods

2.1. Search-based Method

Representative works

Hodosh, M., Young P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell. Research 47, 853–899 (2013)

Hodosh et al. also regard the image caption task as a sorting task. Still, the difference is that the nuclear category correlation analysis technology, is used to project the image and the attribute items extracted from the letters into a public space. **Through training, the image and its corresponding caption have the most remarkable correlation**. Then put the undescribed image in this public space, and select the caption with the highest ranking as the final description by calculating their cosine similarity between all captions.

II. The earlier image caption methods

2.1. Search-based Method

Representative works

Mason, R., Charniak, E.: Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2 (Short Papers) (2014)

Mason and Charniak took the lead in considering the effect of noise on the method. They use visual similarity to search for images from the dataset similar to undescribed images and then obtain captions corresponding to these images. The image captions are ranked by calculating the probability density of words, and finally, the image description with the highest ranking is selected as a result.

II. The earlier image caption methods

2.1. Search-based Method

NOTE

The above methods are all trying to find the description sentence that best matches the query image from the existing image description in the data set. The rationality of this type of method must follow a premise: there must be a caption in the data set that matches the query image. However, in practical applications, this is impossible. Therefore, instead of directly finding the best matching description sentence, some methods try to refine the phrases in the best matching description sentence and synthesize them into a new caption sentence.

II. The earlier image caption methods

2.1. Search-based Method

Such methods rely heavily on existing data sets. Because in the specific image captioning, it reorders the caption of similar images in the data set, and finally uses the description sentence as the description of the query image. Therefore, this method cannot generate new sentences very well, which is also the obvious defect of this method:

- First, if there are only a few particularly good image description sentences in the data set, the final caption will be difficult to produce satisfactory results;
- Secondly, if the query image differs greatly from the image in the data set, for example, the difference in content or style is obvious, it is often difficult to find a similar image in the data set for the query image and it is also difficult to obtain better results.

II. The earlier image caption methods

2.2. Language template-based approaches

This type of method usually **first makes a basic understanding of visual features of the picture and finds out some visual features, such as objects, relationships, attributes and so forth, and then generates captions based on these visual features obtained through a language model**. Generally, multiple sentences are generated to form a candidate set, and then the description sentences in the candidate set are sorted, and the sentence with a higher ranking is selected as the final result. In this type of method, the most important idea is still to extract *handicraft visual features* and then generate captions through a language model.

II. The earlier image caption methods

2.2. Language template-based approaches

Representative works

Yang, Y., et al.: Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. (2011)

Yang et al. proposed an image captioning method that uses the nouns-verbs-scenes-prepositions quadruplet as a sentence template. In order to describe an image, the detection algorithm, is first used to analyze the objects and scenes in the image, and then the language model trained on the ***Gigaword* corpus** is used to determine the verbs, scenes, and prepositions that can be used to form sentences. Then combine the calculated probabilities of all elements and use **Hidden Markov Model** inference to obtain the best quadruplet. Finally, the image caption is generated according to the information of the determined quadruplet and a language template.

II. The earlier image caption methods

2.2. Language template-based approaches

Representative works

Li, S., et al.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning (2011)

Li et al. used the visual model to extract the object, attribute, and spatial relationship information of the image, and defined it in the **triplet of "(attribute-object), preposition, (attribute-object)**." When given a query image, employ web-scale **n-gram** data for phrase selection to collect candidate phrases that may form a triplet. Then, the **dynamic programming method** is used to realize phrase fusion to find the optimal compatible set of phrases as the caption of the query image.

II. The earlier image caption methods

2.2. Language template-based approaches

Representative works

Mitchell, M., et al.: Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (2012)

Mitchell et al. also used **handicraft visual algorithms** to process images. The process of extracting objects, actions, and scenes in the image to represent the image according to the visual algorithm. After that, the entire image caption task is **formulated as the generation of a decision tree**. They cluster and sort object nouns to determine the content to be described, and finally generate the content seen by the computer vision system in detailed descriptive sentences through the **Trigram language model.**

II. The earlier image caption methods

2.2. Language template-based approaches

NOTE

- *This type of method first obtains visual content information from the image, such as objects, attributes, actions, scenes and so forth, and then generates descriptive sentences through the language template. Therefore, this kind of method can usually represent the image visual content elements better, and at the same time, it can also generate grammatically correct captions through the language model.*
- However, sentences generated by language models **are often simple in form and lack diversity**. They **are not natural and fluent in many situations**. Moreover, due to the limitations of language models, their generalization ability needs to be strengthened.

Outline

I. Introduction

II. The earlier image caption methods

III. The recent deep learning methods

IV. Comparison of state-of-the-art methods

V. Experiment conduction result

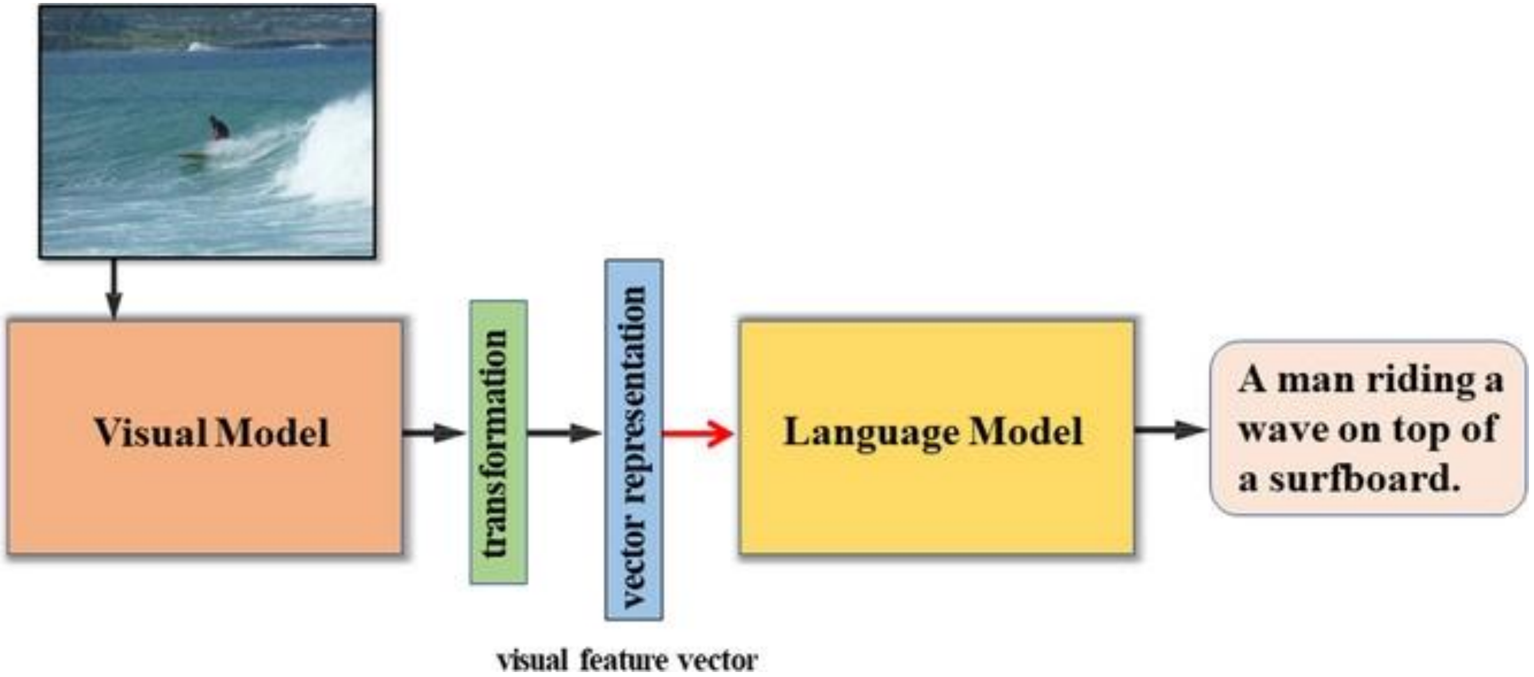
VI. References

III. The recent deep learning methods

With the remarkable progress of deep learning in many fields, more and more researchers have begun to pay attention to neural networks. The same is true in image captioning, especially when the **Encoder-Decoder model** has made significant progress in **machine translation tasks.** Affected by this idea, image captioning has also begun to try this mapping to learn visual features to describe sentences directly from data and outperforms the above two methods.

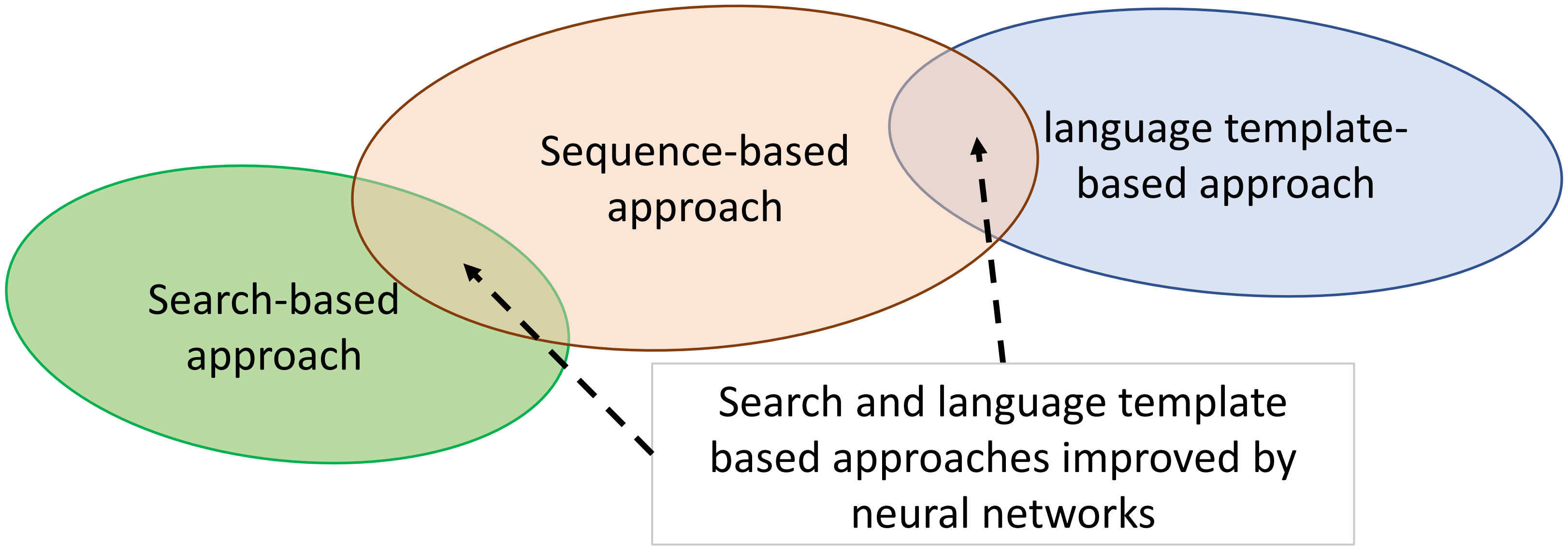
III. The recent deep learning methods

Affected by this idea, image captioning has also begun to try *this mapping to learn visual features to describe sentences directly from data* and outperforms the above two methods. Since this method of image description added to the neural network model mainly uses sequential network models such as **VGG**, **ResNet** and so forth, convolutional neural network **(CNN)** in the encoder or **RNN** and **LSTM** in the decoder part, it is known as the **Sequence-based approach**. The overall framework of the sequence-based approach is illustrated in Figure 3.



III. The recent deep learning methods

The relationship between the **search-based approach, the language template-based approach, and the sequence-based approach** is illustrated in the figure below. Even though deep neural networks are widely adopted for tackling image captioning tasks, different methods may be based on different technical frameworks. Therefore, we classify sequence-based methods into subcategories according to the main technical framework and discuss each subcategory, respectively.



Sequence-based approach

language template-based approach

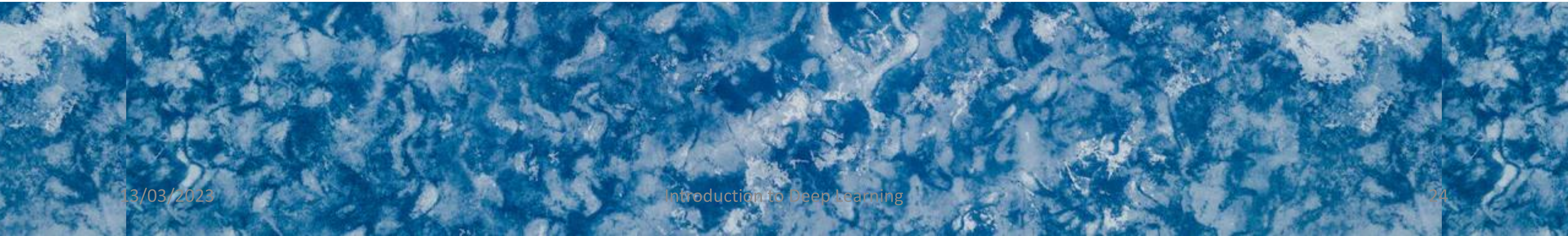
Search-based approach

Search and language template based approaches improved by neural networks

III. The recent deep learning methods

3.1. Search and language template based approaches improved by neural networks

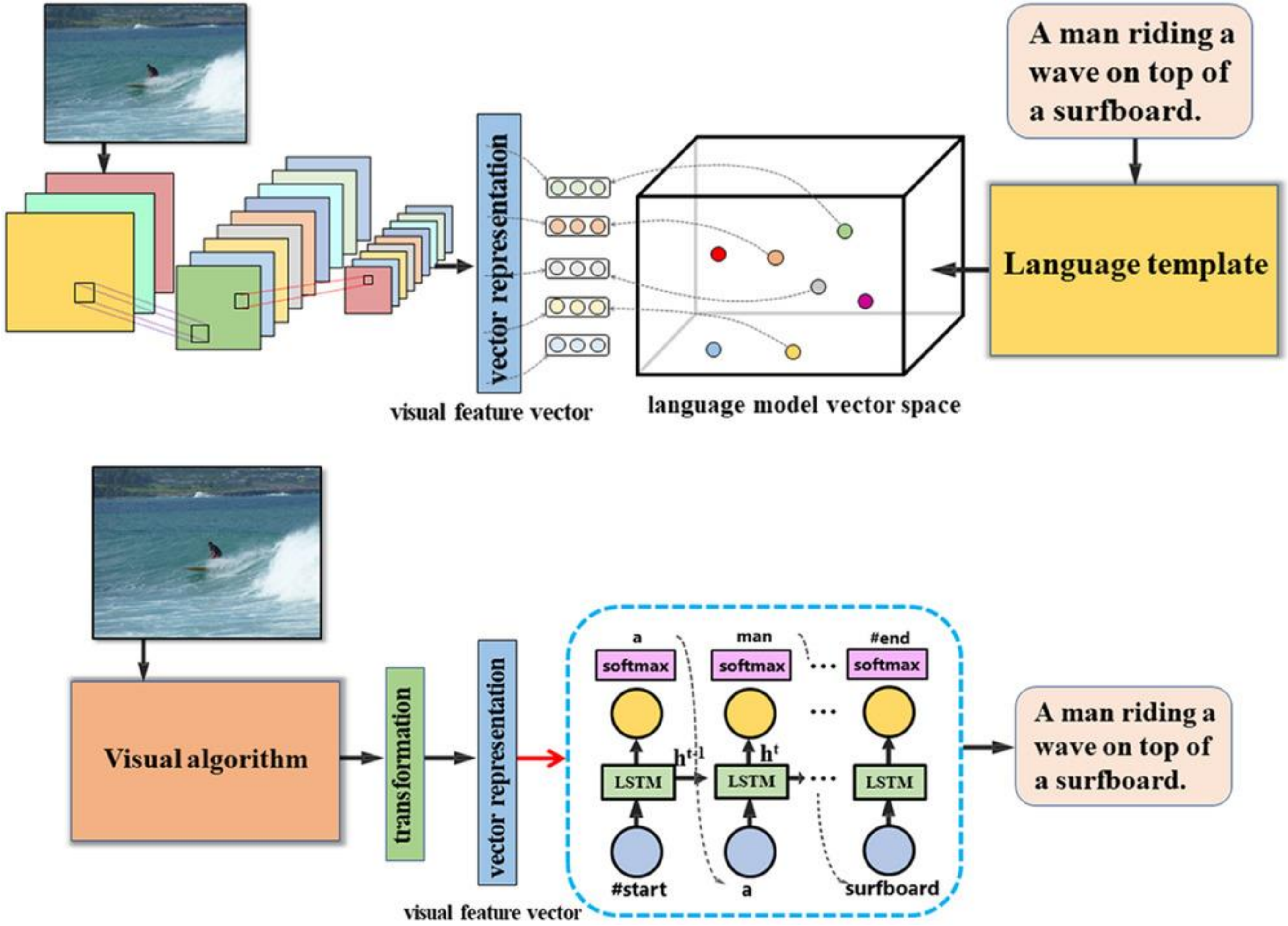
Different from **search-based** and **language template-based** methods, inspired by advances in the field of deep neural networks, deep neural networks are employed to perform image captioning tasks as **an encoder or decoder** part of the **visual-language task**. When the neural network is adopted as an optical encoder, *it mainly learns the expression of images to visual features. In contrast, when adopted as a linguistic decoder, it needs to learn to map transformed visual features from the query image to the corresponding description sentences*



III. The recent deep learning methods

3.1. Search and language template based approaches improved by neural networks

The framework of search and language template based approaches improved by neural networks



A man riding a wave on top of a surfboard.

A man riding a wave on top of a surfboard.

III. The recent deep learning methods

3.1. Search and language template based approaches improved by neural networks

Representative works

Socher, R., et al.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. 2, 207–218 (2014)

Socher et al. employed the CNN model proposed in *"Building high-level features using large scale unsupervised learning. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013)"* to **extract visual features from a query image, analyzed the phrases order and sentence syntax of captions in the data set**, and expressed them as vectors by relying on trees. Then mapped these features to a common vector space through the **maximum margin objective function**. Finally, search for the corresponding image description by calculating the inner product of the image visual features and the description vector.

III. The recent deep learning methods

3.1. Search and language template based approaches improved by neural networks

Representative works

Kiros R., Salakhutdinov R., Zemel R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In: Proceedings of the NIPS Workshop on International MachineLearningSociety (2014)

Kiro et al. learned an image-caption vector space and a language model to decode this space. The author uses **CNN and LSTM models** to encode image and caption sentences respectively. The encoded image and sentence representations are mapped to the same computing space through two **fully connected networks**, and then the CNN model and the LSTM model are trained separately. Besides, the author proposes Log-bilinear neural language models and Multiplicative neural language models to decode the representation vector and form a new description.

III. The recent deep learning methods

3.1. Search and language template based approaches improved by neural networks

NOTE

In the above studies, although manual visual algorithms, language models, or measurement systems are still used in the entire framework, the performance has been improved due to adopted neural network models. **However, the framework formed by the manual algorithms and the neural networks trained by separation often does not achieve optimal performance**. There are roughly three reasons for this:

1. Multiple modules in the entire framework cannot learn from each other during the design or training process;
2. The objective training function deviates from the overall performance index of the system;
3. The algorithm itself cannot completely exclude the limitations of artificial design methods, and it imposes restrictions on the generated descriptions.

III. The recent deep learning methods

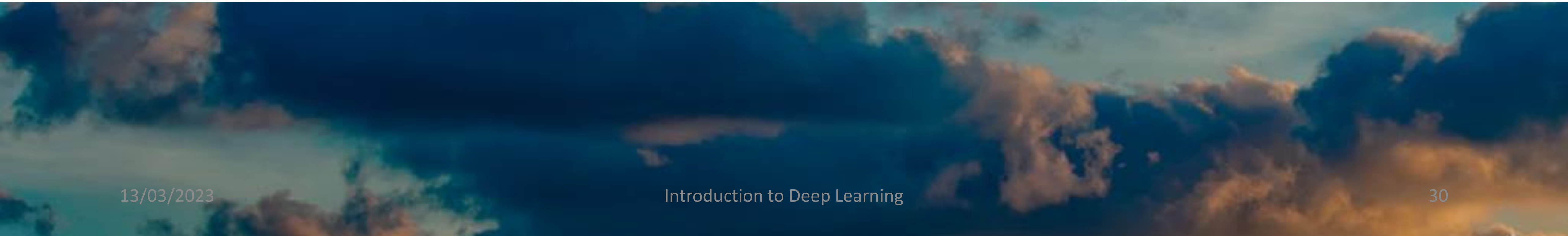
3.1. Search and language template based approaches improved by neural networks

Therefore, researchers try to generate descriptions through **end-to-end systems**. They hope that by *reducing manual pre-processing and subsequent processing*, as much as possible to make the model from the original input to the final output, using **a pipelined model**, to avoid the inherent shortcomings of the multi-module mentioned above. Moreover, the *end-to-end approach reduces the project's complexity and gives the model more room for free play*.

III. The recent deep learning methods

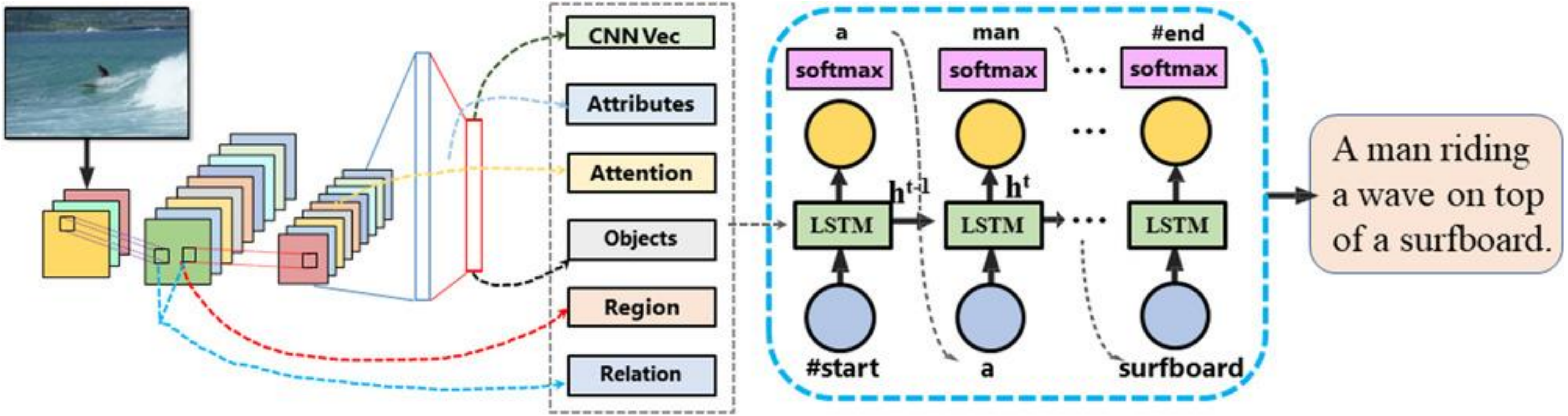
3.2. Image captioning with high-level representations

A glance at the image is enough for humans to point out and describe many details about the visual scene. However, it turns out that this extraordinary ability is an elusive task for our visual recognition model. Some studies expect to design a sufficiently rich model to simultaneously reason about the high-level semantic contents of the query image and its representation in the field of natural language.



III. The recent deep learning methods

3.2. Image captioning with high-level representations



III. The recent deep learning methods

3.2. Image captioning with high-level representations

Representative works

Wu, Q., Shen, C. & Liu, L.: What value do explicit high level concepts have in vision to language problems? In: IEEE conference on computer vision and pattern recognition, pp. 203-212 (2016)

Wu et al. provided a new idea. The author believes that the feature map after the CNN model should not be directly connected to the image caption problem, but should have high-level semantic features. **They extracted the 256 words that appear most frequently (at least 5 times) from the descriptive sentences** in the training set as the most representative attributes. The VGG network model pre-trained on ImageNet is used to modify its final output into a 256-dimensional attribute vector, corresponding to the extracted 256 attributes, and this CNN structure is used as the encoder.

III. The recent deep learning methods

3.2. Image captioning with high-level representations

Representative works

Wu, Q., Shen, C. & Liu, L.: What value do explicit high level concepts have in vision to language problems? In: IEEE conference on computer vision and pattern recognition, pp. 203-212 (2016)

Since a picture may correspond to multiple attributes, through the training of multi-label task classification, the features extracted by the CNN structure contain semantic information. Furthermore, to ensure that the model can effectively extract semantic information, the author also added a detector to further improve the accuracy of the model. The decoder continues to use the LSTM structure, and finally generates an image caption rich in semantic information.

III. The recent deep learning methods

3.2. Image captioning with high-level representations

Representative works

Yao, T., et al.: Exploring visual relationship for image captioning. In: Proceedings of the European Conference On Computer Vision (ECCV) (2018)

Yao et al. proposed a **GCN-LSTM** structure to describe the relationship between objects under the framework of the attention mechanism. They propose the structure of the combination of **graph convolutional network (GCN) and LSTM to integrate the semantics and the relationship of objects in space into the image encoder**. The relational graph is constructed according to the spatial and semantic connections of the objects detected in the image. Then, the relationship graph refines the representation of each region through the GCN graphic structure to obtain the regional-level relationship perception features, and finally inject it into the LSTM to generate a description sentence. Then, the representation of each region in the relationship graph is refined through the GCN graph, to transform the regional-level relationship perception feature, and finally injected into the LSTM to generate a caption.

III. The recent deep learning methods

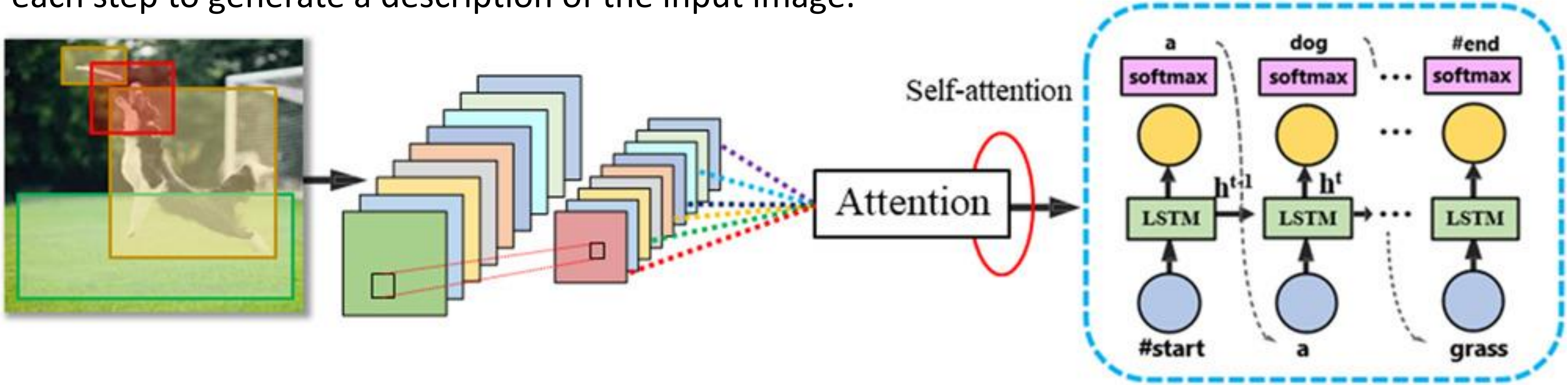
3.3. Enhanced image captioning with attention correction

Images can convey rich semantics due to rich visual information. However, only the most prominent content needs to be paid attention to in image captioning tasks. Moreover, while the neural network exhibits powerful representation capabilities, it also brings redundant and disturbing information due to its complex structure.

III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction

Inspired by the human visual attention mechanism, methods of using attention to guide image caption generation are proposed. In such methods, the **attention mechanism** based on various kinds of feature representation of the input image is integrated into the language generation framework to make the generation process focus on the interest regions of the visual features at each step to generate a description of the input image.



III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction

Representative works

Vinyals, O., et al.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

Vinyals et al. [49] proposed a model based on CNN+LSTM. This structure provides a general idea for image description tasks. The framework first encodes the image with GoogleNet [50] as feature vectors, and then decodes them with LSTM, which outputs the probability of all words in the word list, selects the highest word as the output, and finally forms the image description. The model achieved state-of-the-art performance at the time.

III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction

Representative works

Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (2015)

Xu et al. added an attention mechanism on this basis to further improve the performance of the model. The author also uses the image as input, **CNN as the encoder to extract the image features** to form a feature map, and then **through the attention mechanism to enhance or suppress the feature map**. As the input data into the LSTM model, the data after the attention mechanism at different moments will be adjusted by the output data of the LSTM model at the previous moment, and finally, the image description is generated through the LSTM model.

III. The recent deep learning methods

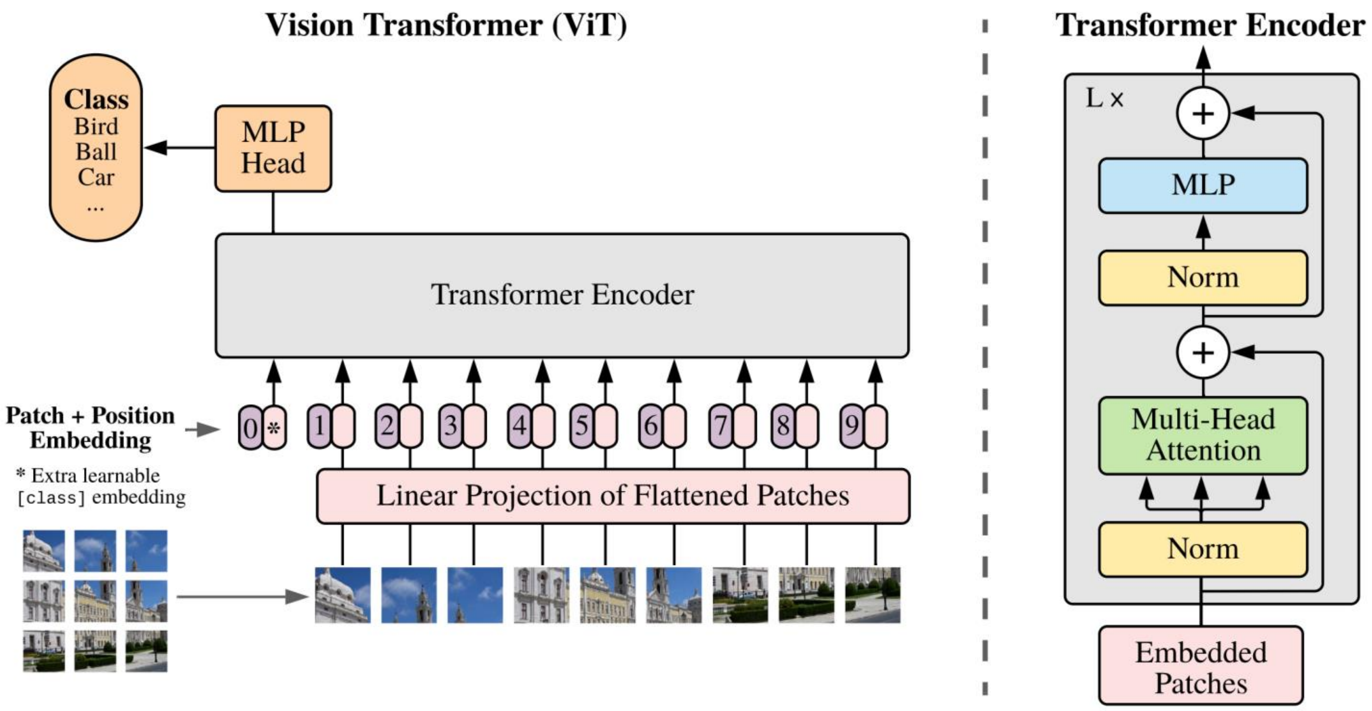
3.3. Enhanced image captioning with attention correction

Recently, transformers have shown good performance when dealing with serialized information. More importantly, the transformer has been recognized as the latest technology for sequence modeling tasks such as language understanding and machine translation. Some studies have adopted the newest transformer architecture for image captioning.

III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction

Transformer-like architectures could be applied directly on visual context patches, thus excluding or limiting the usage of the convolutional operator. On this line, *Liu et al.* designed the first non-convolutional architecture for image captioning. Specifically, a pre-trained **Vision Transformer** in is employed as an encoder, and then a general Transformer is accepted as decoder to generate captions.



III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction

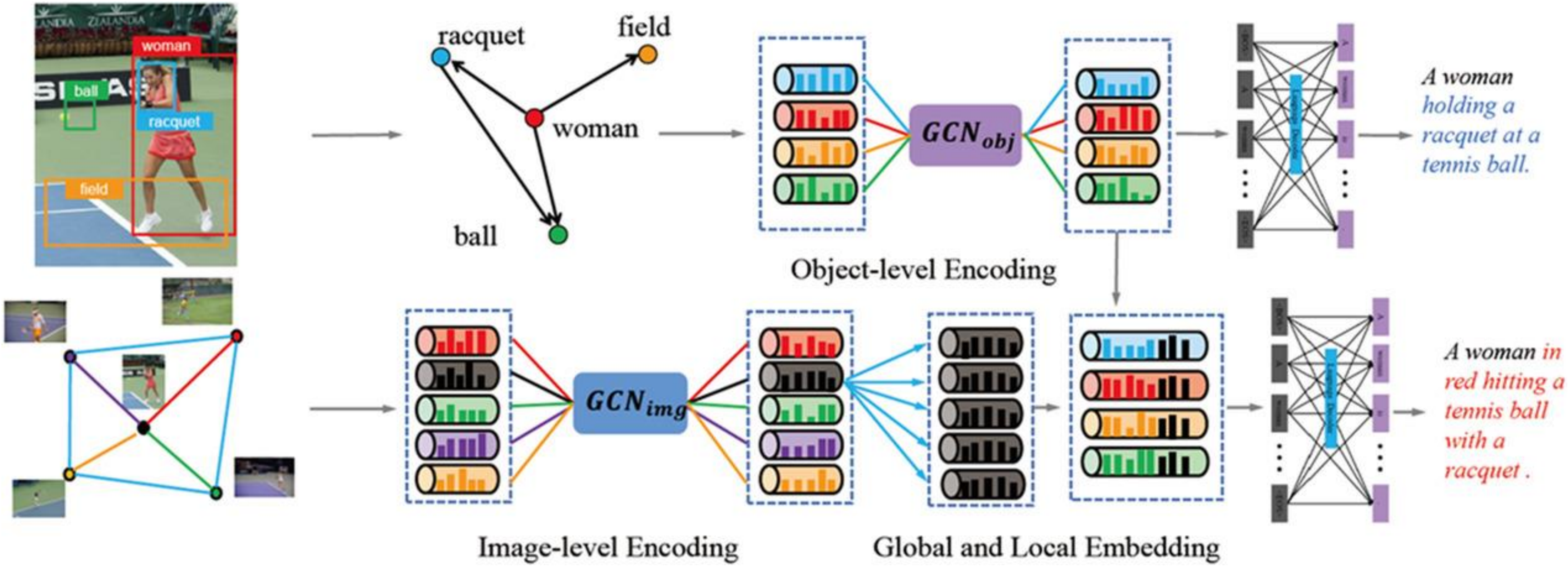
Representative works

Dong, X., Long, C., Xu, W., et al.: Dual graph convolutional networks with transformer and curriculum learning for image captioning. arXiv:2108.02366 (2021)

Dong et al. proposed dual **graph convolutional networks (Dual-GCN)** with **transformer** and **curriculum learning** to explore the contextual relevance between contextual images for image captioning. Two independent GCNs encode the entire image and the objects from the image, and then the captions are generated by a Transformer linguistic decoder. *Ji et al.* introduce a **Global Enhanced Transformer** to extract a more comprehensive global representation and then adaptively guide the decoder to generate high-quality captions.

III. The recent deep learning methods

3.3. Enhanced image captioning with attention correction


woman
ball
racquet
field
racquet
field
woman
ball
GCN_{obj}
Object-level Encoding
A woman holding a racquet at a tennis ball.
GCN_{img}
Image-level Encoding
Global and Local Embedding
A woman in red hitting a tennis ball with a racquet .

Outline

I. Introduction

II. The earlier image caption methods

III. The recent deep learning methods

IV. Comparison of state-of-the-art methods

V. Experiment conduction result

VI. References

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

The image captioning task is comprehensive of computer vision and NLP. It can be simply understood that this task requires the model to recognize objects, actions, scenes, and relationships between objects in the image, and then map these contained visual contents into descriptive sentences. In general, this vision-language task requires two basic requirements:

1. **The correctness of the grammar**—the language grammar needs to be followed during the mapping process to make the result readable;
2. **The richness of the description sentence**—the generated caption needs to be able to accurately describe the details of the corresponding image, and produce a sufficiently complex description.

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

- **The Bilingual Evaluation Understudy (BLEU)** method is adopted to evaluate the quality of translated sentences in machine translation. It compares each translation segment with a set of reference translations with good translation quality and calculates each segment score then estimates the overall quality of the translation.

- Bleu Scores are between 0 and 1. A score of **0.6 or 0.7** is considered the best you can achieve. Even two humans would likely come up with different sentence variants for a problem, and would rarely achieve a perfect match. For this reason, a score closer to 1 is unrealistic in practice and should raise a flag that your model is overfitting.

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

N-gram

- An ‘n-gram’ is actually a widely used concept from regular text processing and is not specific to NLP or Bleu Score. It is just a fancy way of describing “a set of ‘n’ consecutive words in a sentence”.
- For instance, in the sentence “The ball is blue”, we could have n-grams such as:
  - 1-gram (unigram): “The”, “ball”, “is”, “blue”
  - 2-gram (bigram): “The ball”, “ball is”, “is blue”
  - 3-gram (trigram): “The ball is”, “ball is blue”
  - 4-gram: “The ball is blue”

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Precision

This metric measures the number of words in the *Predicted Sentence* that also occur in the *Target Sentence*.

Let's say, that we have:

- **Target Sentence**: "He eats an apple"
- **Predicted Sentence**: "He ate an apple"

→ *Precision = 3 / 4*

We would normally compute the Precision using the formula:

Precision = Number of correct predicted words / Number of total predicted words

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Repetition

The first issue is that this formula allows us to cheat. We could predict a sentence:

- **Target Sentence**: He eats an apple
- **Predicted Sentence**: He He He

and get a perfect Precision = 3 / 3 = 1

Precision = 3 / 3 = 1

Precision = Number of correct predicted words / Number of total predicted words

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Multiple Target Sentences

Secondly, as we’ve already discussed, there are many correct ways to express the same sentence. In many NLP models, we might be given multiple acceptable target sentences that capture these different variations.

We account for these two scenarios using a modified Precision formula which we’ll call “**Clipped Precision**”.

IV. Comparison of state-of-the-art methods

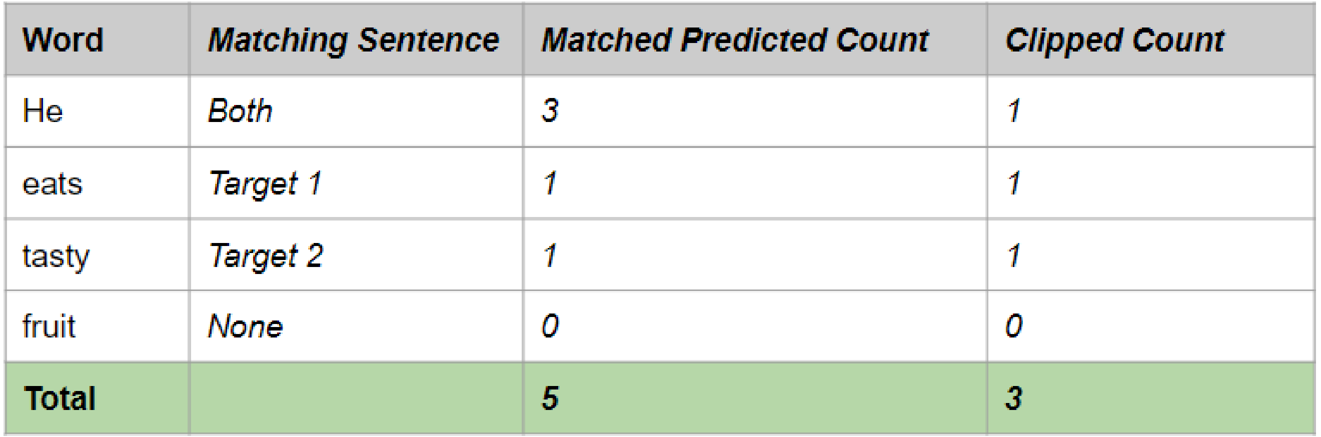
4.1. Evaluation metrics

BLEU Score

Clipped Precision

Let’s say, that we have the following sentences:

- **Target Sentence 1**: He eats a sweet apple
- **Target Sentence 2**: He is eating a tasty apple
- **Predicted Sentence**: He He He eats tasty fruit



Word	*Matching Sentence*	*Matched Predicted Count*	*Clipped Count*
He	*Both*	*3*	*1*
eats	*Target 1*	*1*	*1*
tasty	*Target 2*	*1*	*1*
fruit	*None*	*0*	*0*
Total		***5***	***3***

We now do two things differently:

1. We compare each word from the predicted sentence with all of the target sentences. If the word matches any target sentence, it is considered to be correct.
2. *We limit the count for each correct word to the maximum number of times that that word occurs in the Target Sentence*.

IV. Comparison of state-of-the-art methods

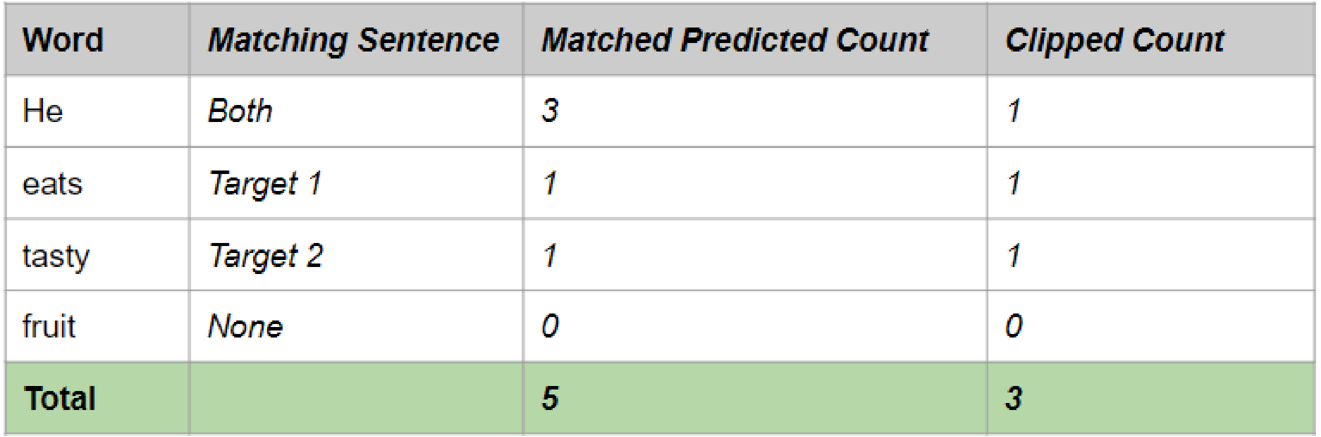
4.1. Evaluation metrics

BLEU Score

Clipped Precision

Let’s say, that we have the following sentences:

- **Target Sentence 1**: He eats a sweet apple
- **Target Sentence 2**: He is eating a tasty apple
- **Predicted Sentence**: He He He eats tasty fruit



Word	*Matching Sentence*	*Matched Predicted Count*	*Clipped Count*
He	*Both*	*3*	*1*
eats	*Target 1*	*1*	*1*
tasty	*Target 2*	*1*	*1*
fruit	*None*	*0*	*0*
Total		***5***	***3***

Clipped Precision = Clipped number of correct predicted words / Number of total predicted words

Clipped Precision = 3 / 6

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

How is Bleu Score calculated?

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

The first step is to compute Precision scores for 1-grams through 4-grams.

IV. Comparison of state-of-the-art methods

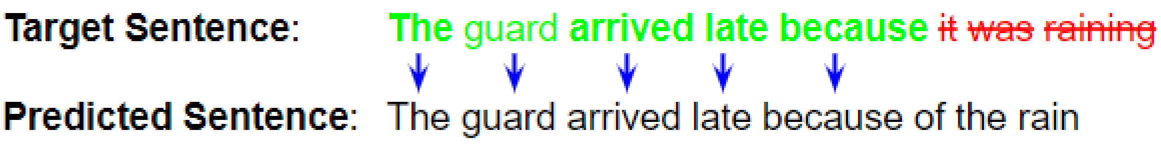
4.1. Evaluation metrics

BLEU Score

Precision 1-gram

- We use the Clipped Precision method that we just discussed.

Precision 1-gram = Number of correct predicted 1-grams / Number of total predicted 1-grams



Target Sentence: The guard arrived late because ~~it was raining~~

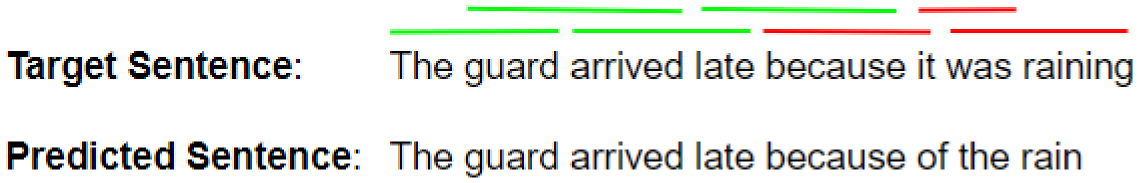
Predicted Sentence: The guard arrived late because of the rain

So, Precision 1-gram (p_1) = 5 / 8

Precision 2-gram

- Let’s look at all the 2-grams in our predicted sentence:

Precision 2-gram = Number of correct predicted 2-grams / Number of total predicted 2-grams



Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

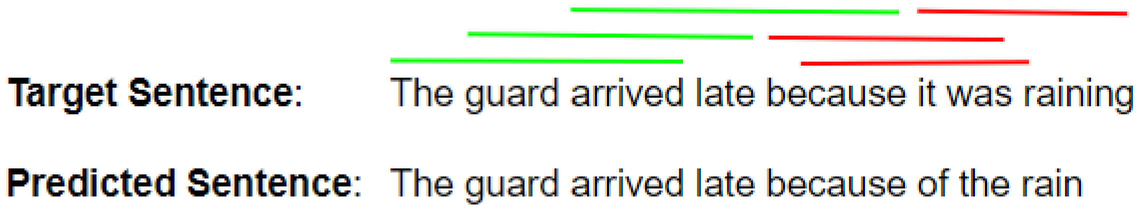
So, Precision 2-gram ($p2$) = 4 / 7

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Precision 3-gram

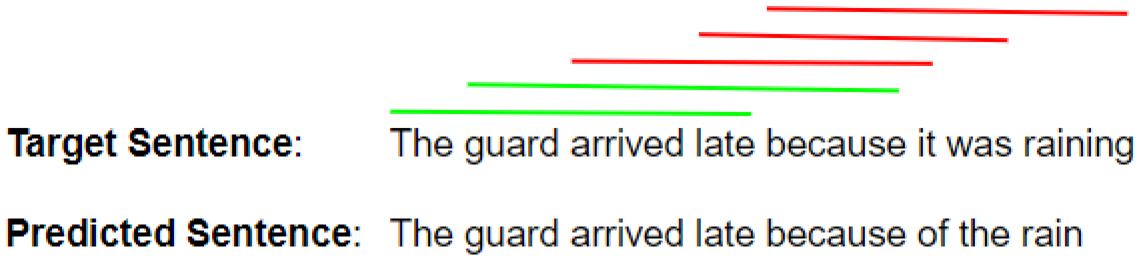


Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

Similarly, Precision 3-gram (p_3) = 3 / 6

Precision 4-gram



Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

And, Precision 4-gram (p4) = 2 / 5

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Geometric Average Precision Scores

Next, we combine these Precision Scores using the formula below. This can be computed for different values of N and using different weight values. Typically, we use *N = 4* and uniform weights *wn = N / 4*

$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} {p_n}^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

- (*p1*) = 5 / 8
- (*p2*) = 4 / 7
- (*p3*) = 3 / 6
- (p4) = 2 / 5

GAP = 0,517

Target Sentence: The guard arrived late because it was raining
Predicted Sentence: The guard arrived late because of the rain

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

Brevity Penalty

If you notice how Precision is calculated, we could have output a predicted sentence consisting of a single word like “The’ or “late”. For this, the 1-gram Precision would have been 1/1 = 1, indicating a perfect score.

This is obviously misleading because it encourages the model to output fewer words and get a high score. To offset this, the ***Brevity Penalty penalizes sentences that are too short***.

$$Brevity\ Penalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}$$

- ***r*** is *target length = number of words in the target sentence* and
- ***c*** is *predicted length = number of words in the predicted sentence*

Precision = Number of correct predicted words / Number of total predicted words

Clipped Precision = Clipped number of correct predicted words / Number of total predicted words

IV. Comparison of state-of-the-art methods

4.1. Evaluation metrics

BLEU Score

BLEU Score

Finally, to calculate the Bleu Score, we multiply the Brevity Penalty with the Geometric Average of the Precision Scores.

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

Bleu Score can be computed for different values of N. Typically, we use N = 4.

- BLEU-1 uses the unigram Precision score
- BLEU-2 uses the geometric average of unigram and bigram precision
- BLEU-3 uses the geometric average of unigram, bigram, and trigram precision
- and so on.

IV. Comparison of state-of-the-art methods

4.2. Image caption datasets

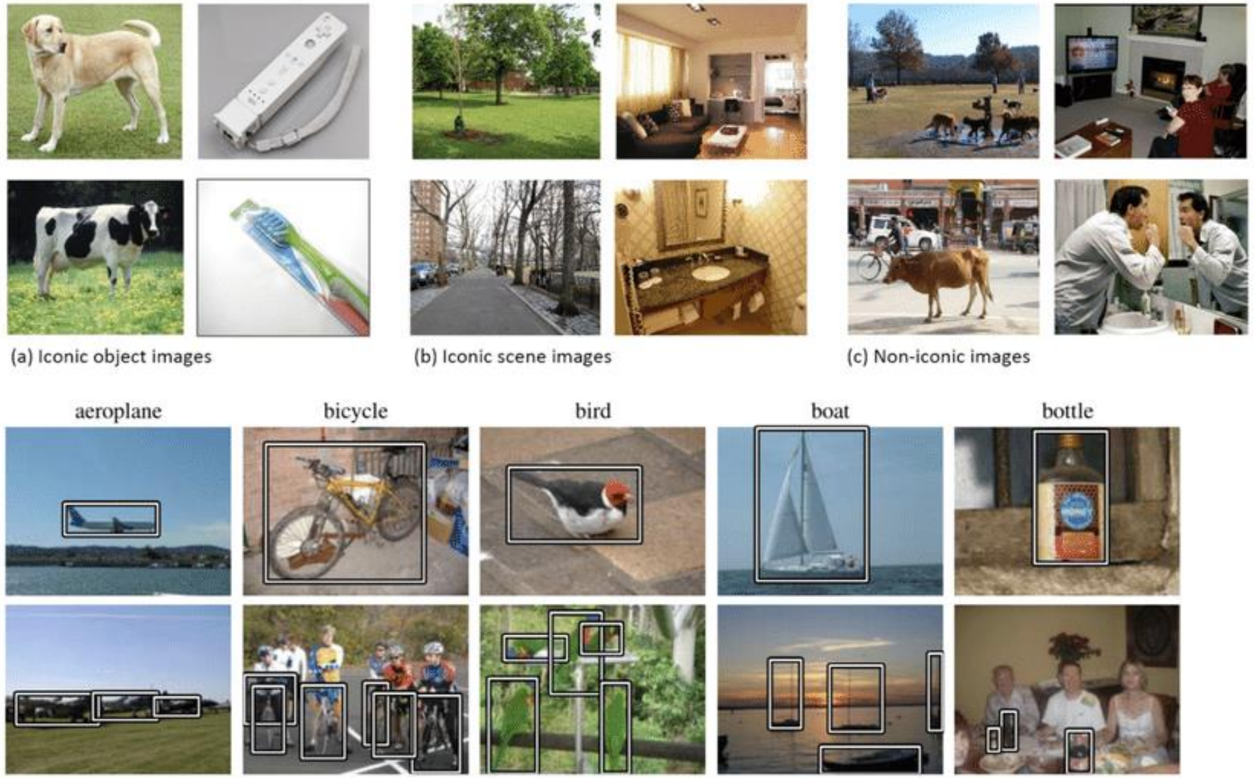
So far, there have been a large number of data sets used for image captioning. These data sets are different to a certain extent in terms of data collection and sorting, presentation of data labels, as well as the volume and specifications of the datasets, which lays the data foundation for the task of image description generation

IV. Comparison of state-of-the-art methods

4.2. Image caption datasets

- **MS COCO dataset**

A large-scale dataset launched by Microsoft in 2014 that can be used for tasks such as image recognition, object detection, semantic segmentation, and image caption. The images in the dataset consist of nearly 100 object categories from images of daily complex scenes containing ordinary objects under natural backgrounds, and each image is artificially annotated using **Amazon Mechanical Turk (AMT)**.



(a) Iconic object images

(b) Iconic scene images

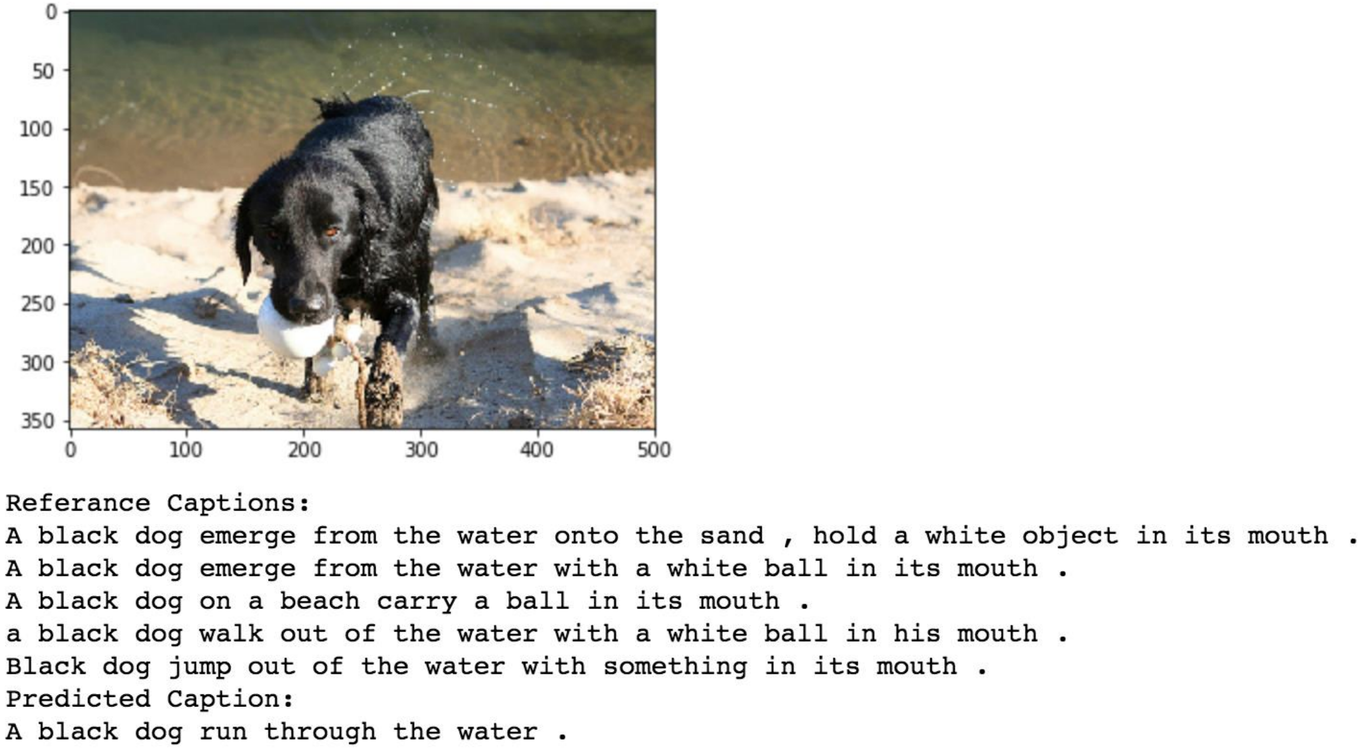
(c) Non-iconic images

IV. Comparison of state-of-the-art methods

4.2. Image caption datasets

- **Flickr8K dataset**

Released for public use by researchers in 2013. The images in the dataset are all from the photo and image sharing website **Flickr** and contain 8000 images. Compared with MS MSCOCO, the data scale is small, and the image content is mainly human and animal. The label caption is also through crowdsourcing services by Amazon's manual labeling platform. Each image has five sentences as description.



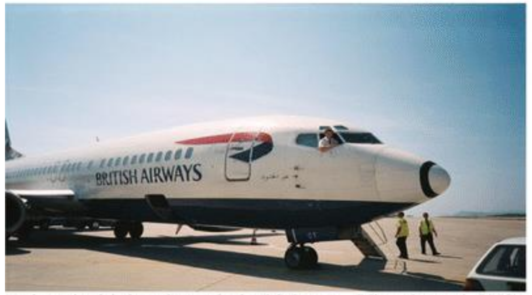
```
Referance Captions:
A black dog emerge from the water onto the sand , hold a white object in its mouth .
A black dog emerge from the water with a white ball in its mouth .
A black dog on a beach carry a ball in its mouth .
a black dog walk out of the water with a white ball in his mouth .
Black dog jump out of the water with something in its mouth .
Predicted Caption:
A black dog run through the water .
```

IV. Comparison of state-of-the-art methods

4.2. Image caption datasets

- **Flickr30K dataset**

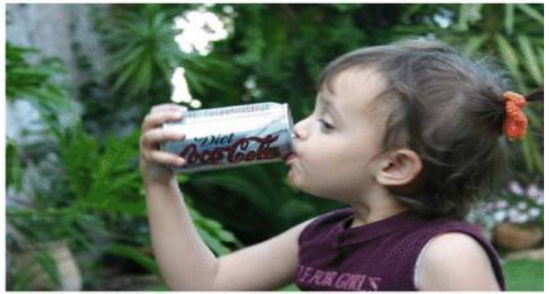
An extension of Flickr8K, contains 31783 image data, each image has five sentences corresponding description.



a plane is driving through the highway. (Vinyals et al. [23])

a plane with a man hanging out of the cockpit. (Kiros et al. [20])

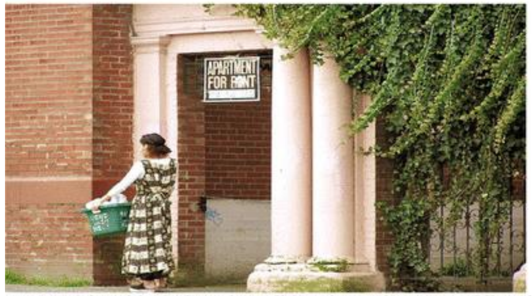
Proposed: a British Airways plane on the airstrip.



a young girl is holding a large bottle in her arms. (Vinyals et al. [23]):

a little girl in a purple shirt is drinking. (Kiros et al. [20])

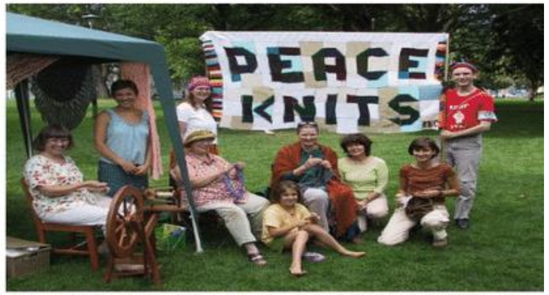
Proposed: a young girl is taking a drink from a can of diet coca cola.



a woman in a black dress is walking down the street. (Vinyals et al. [23])

a woman holding a clothes basket. (Kiros et al. [20])

Proposed: a woman standing in front of a building with an apartment for rent.



a crowd of people sitting on a lawn holding a sign. (Vinyals et al. [23])

a group of young and older people gathered at garden. (Kiros et al. [20])

Proposed: people are knitting next to a Peace Knits sign

References

1. Gaifang Luo, Lijun Cheng, Chao Jing, Can Zhao, Guozhu Song: A thorough review of models, evaluation metrics, and datasets on image captioning. (2021)
2. Foundations of NLP Explained — Bleu Score and WER Metrics | Ketan Doshi Blog (ketanhdoshi.github.io)

thank you