Shanghai Jiao Tong University
AI3611

Image Caption

Yan Han (519030910404)

June 12, 2022

1 Introduction

Image caption, whose goal is to automatically generate a natural language description of an image, has recently received lots of attention in Computer Vision. While covering both Computer Vision and Natural Language Processing, it can be regarded as a cross-modal task. In the following, 1.1 introduces the different ways to process various modal data and the whole framework of the Encoder-Decoder model. 1.2 shows various metrics that measure the performance of the model。1.3 introduces a method to mitigate the effects of noise caused by cumulative errors, which is known as schedule sampling [2]. The source code is available at `https://github.com/wolfball/Intellisense-Cognitive-Practice`.

1.1 Pipeline

The pipeline is based on [9], which is an Encoder-Decoder framwork (See Fig.1). For images, we use an encoder to get their features and for natural language data, we use word embeddings to represent. Then, by combining image features with the current timestep word embedding, we can use a decoder to produce the next word. For encoder, we use ResNet to extract a 14×14 feature map from the input image. Then we employ attention mechanism to get more focused feature. For decoder, we utilize LSTM to produce image caption (See Figure 2).

1.2 Metrics

To comprehensively measure the performance of the model, we use Bleu [7], Rouge [5], METEOR [4], CIDEr [8], SPICE [1], SPIDEr [6] to evaluate the predictions. For all metrics, the higher the better.

Bleu Bilingual Evaluation Understudy, which is used to analyze the degree of co-occurrence of n-tuples in candidate translation and reference translation. The advantage of BLEU is that it considers n-gram granularity rather than words, allowing for longer matching information; The disadvantage is that no matter what N-gram is matched, it is treated the same thus being disturbed by common words (like ”is”, ”the”).

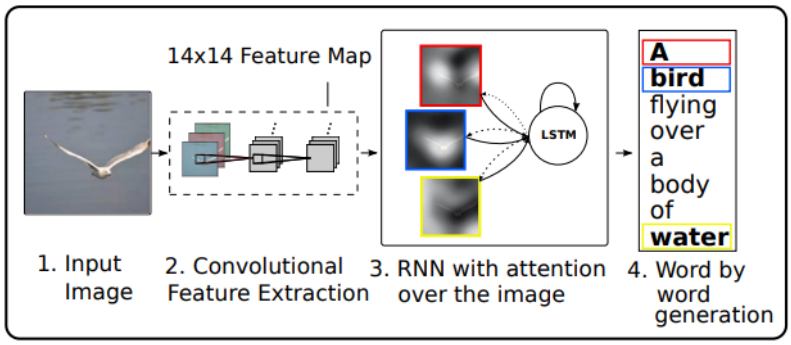


Figure 1: pipeline

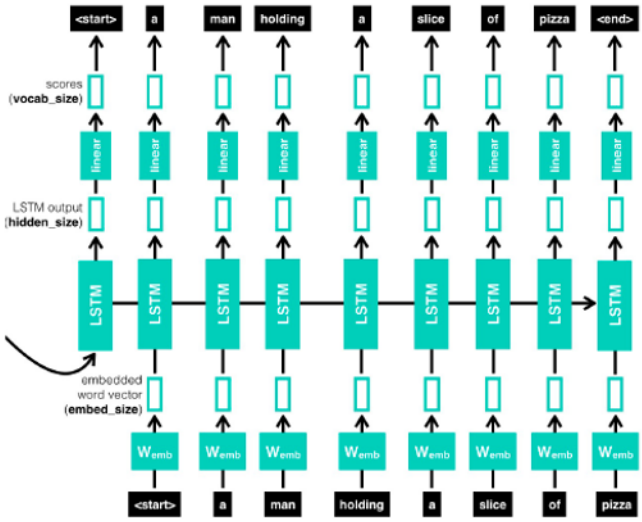


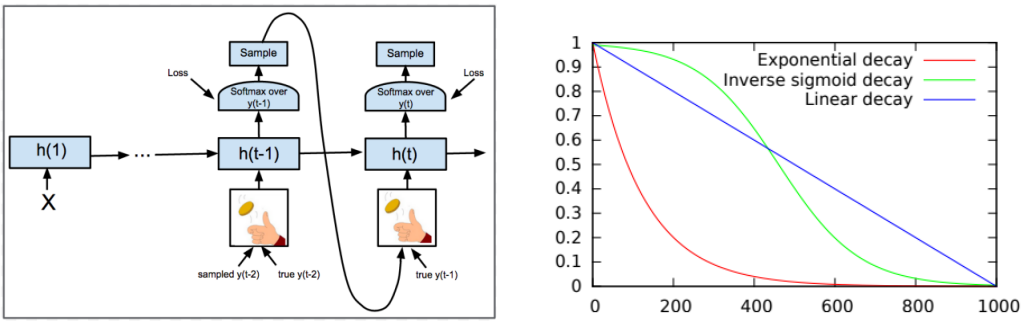
Figure 2: pipeline

Rouge ROUGE is calculated based on the recall rate, so it is the evaluation standard for automatic summary tasks.

METEOR METEOR is an improvement based on Bleu. Using WordNet to calculate the matching relationships between specific sequences, synonyms, roots and affixes, and definitions improves BLEU's performance and makes it more relevant to manual discriminations.

CIDEr CIDEr is specifically designed for image labeling problems. It regards each sentence as a "Document" and expresses it in the form of Term Frequency Inverse Document Frequency (TF-IDF) vector. Through weight calculation of each n-tuple (TF-IDF), Cosine similarity of reference caption and model generated caption, CIDEr can measure the consistency of image annotation. IDF provides a way to measure the salience of words by discounting the importance of words that are easy to come up with, but don't contribute much to visual content information.

SPICE Semantic Propositional Image Caption Evaluation was also specifically designed for the image Caption problem. The first four methods were all based on N-gram calculations, so SPICE was designed to solve this problem. SPICE uses graph-based semantic representations to encode objects, attributes, and relationships with caption. It first parsed the evaluated caption and reference captions into Syntactic dependencies using



(a) Illustration of the Scheduled Sampling approach

(b) Decay schedules

Figure 3: schedule sampling

Probabilistic Context-Free Grammar (PCFG) dependency Parser Trees, and then map dependency Tree to Scene Graphs using a rules-based approach. Finally, calculate the F-score of objects, attributes and relationships in the caption to be evaluated.

SPIDEr SPIDEr is a linear combination of SPICE and CIDEr .The reason optimizing SPICE gives poor captions is that SPICE ignores syntactic quality. More generally, we argue that a good image captioning metric should satisfy two criteria: (1) captions that are considered good by humans should achieve high scores; and (2) captions that achieve high scores should be considered good by humans. SPICE satisfies criterion 1, but not criterion 2. This metric automatically satisfies criterion 1, since both SPICE and CIDEr do. Also, when we optimize for SPIDEr, it satisfies criterion 2 to a much greater degree than existing metrics [6].

1.3 Schedule Sampling

While training, we usually use the ground truth to predict next word. However, during the testing stage, we can only use the previous predictions to produce the next one. That is to say, if the prediction in time t is error, the following predictions are all built on the wrong words, thus resulting in accumulated error. In order to ease this negative impact, [2] introduce a new mathod called **schedule sampling**.

During training, the network will no longer completely adopt the real sequence markers as the next input, but selects the real markers with a probability p and the output of the model itself with $1 - p$ (See Fig.3(a)). The size of p varies during training, as does the learning rate. The idea is that the network training is not sufficient at the beginning, so p should be selected as large as possible, that is, the ground true should be used as far as possible. Then, as the training goes on, the model training becomes more and more sufficient, and p also decreases, that is, the output of the model should be selected as far as possible. In this way, the model training and prediction are as consistent as possible (See some examples in Fig.3(b)).

2 Experiments and results

This section mainly focuses on the experiments of baseline and schedule sampling. 2.1 describes the experiment in detail. 2.2 shows the best and the worst, objective and subjective evaluations for each metric.

2.1 Experimental Detail

We select Flicker8k as our data set. For pre-processing, we resize all images to 256×256, and randomly apply horizontal flip with probability of 0.5, and normalize them such that their means are (0.485, 0.456, 0.406) for each channel and standard variances are (0.229, 0.224, 0.225) for each channel. For each batch with the size of 128, we use zero padding to pad captions to the max length of the data set. During training, we use Resnet to encode each image to the size of (14, 14, 2048), where 2048 is the channel number. Then we utilize attention mechanism to generate a feature vector with dimension of 2048. Concatenated with current word embedding, this feature vector is feed into LSTM network to predict the next word embedding. In baseline configuration, the embedding dimension is 300, attention dimension(size of the attention network) is 256 and decoder size(size of decoder's RNN) is 256. Finally, cross entropy loss between prediction and truth captions is computed. For optimizer, we employ RMSprop with initial learning rate as 0.001. All experiments are executed within 45 epochs.

For schedule sampling, we try three decay schedules parameterized by k:

$$\text{Linear decay:} \quad p = \max(0.1, 1 - kx) \tag{1}$$

$$\text{Exponential decay:} \quad p = k^x \tag{2}$$

$$\text{Inverse sigmoid decay:} \quad p = \frac{k}{k + e^{\frac{x}{k}}} \tag{3}$$

where x is the current epoch. All experiments are ran with 45 epochs.

The results are in Table 1. All experiments with schedule sampling outperform the baseline. In linear decay, $k = \frac{1}{90}$ performs the best; In exponential decay, the performance increases when k increases; In inverse sigmoid decay, the performance has the same trend with exponential decay.

2.2 Objective and Subjective Evaluation

The generated best and worst indicator samples are shown in Figure 4. The followings are some comparisons.

Bleu Take Bleu-4 for example, the best sample is the first image on the left, scoring 1.0, whose prediction is similar to the references. The worst sample is the first image on the right, scoring 0.0, whose prediction has many grammar mistakes.

Rouge The best sample is the second image on the left, scoring 0.880, whose prediction has semantic similarity but not fully correct. The worst sample is the second prediction on the right, scoring 0.102, whose prediction has obvious mistakes.

METEOR The best sample is the first image on the left, scoring 0.551, whose prediction is satisfying. The worst sample is the third prediction on the left, scoring 0.021, whose prediction does not recognize "women" but "group of people", and has a grammar mistake.

CIDEr The best sample is the first image on the left, scoring 3.912, whose prediction is satisfying. The worst sample is the third prediction on the right, scoring 0.0, whose prediction has obvious mistakes.

SPICE The best sample is the fourth image on the left, scoring 0.636, whose prediction has semantic similarity but recognize wrong color. The worst sample is the fourth prediction on the right, scoring 0.0, whose prediction has obvious mistakes.

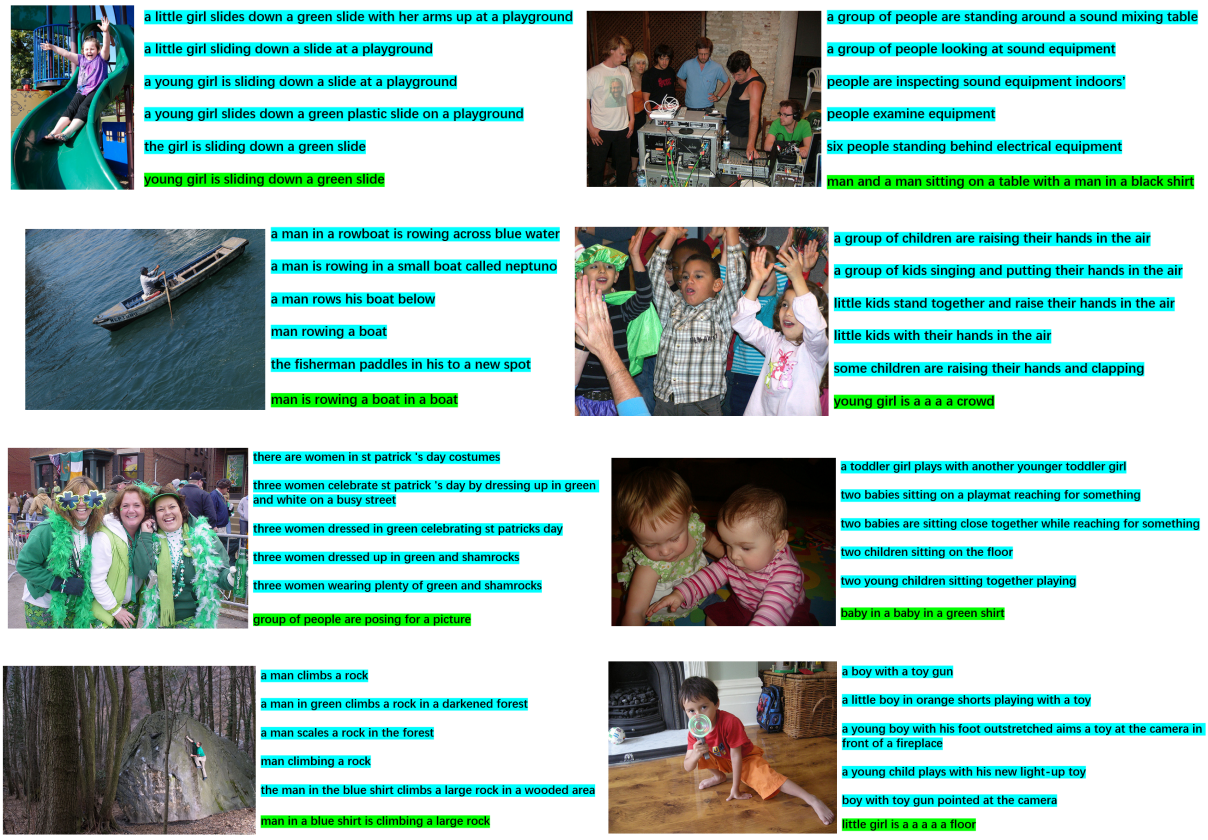


Figure 4: Visualization of the results of image caption. Here we shows 8 samples. For each sample, the image is on the left and the captions are on the right where the blue captions are the references and the green one is the prediction.

SPICEr The best sample is the first image on the left, scoring 2.234, whose prediction is satisfying. The worst sample is the fourth prediction on the right, scoring 0.0, whose prediction has obvious mistakes.

3 Sinusoidal Decay

Inspired by [3], we propose a sinusoidal decay which is different from the above three decay methods. The decay schedule is shown in Figure 5 and the formula is:

$$p = \frac{1}{2}(1 + \cos(\frac{2\pi x}{k})) \tag{4}$$

Intuitively, p with high value means learning more from ground truth and low value means practising by model itself. Thus, sinusoidal decay can help model to learn and practise over and over again, which repeatedly improves the performance.

We have tried $k \in \{5, 9, 15, 20\}$, i.e. different periods. And the result is in Table 1 where shows that $k = 5$ performs the best.

4 Conclusion

In summary, we describe the basic structure for image caption and the methods to dealing with muti-modal data. Plus, we realise the schedule sampling with performs better than baseline. For each metric, we have printed the best and the worst example with objective and subjective evaluation.

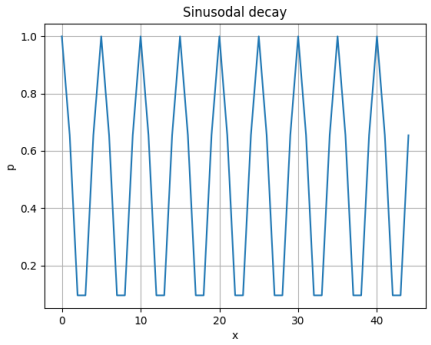


Figure 5: Sinusoidal decay with $k = 5$.

Moreover, we present a new sinusoidal decay which outperforms the original three decay schedules. The best SPIDEr is 0.334 with sinusoidal schedule sampling (k=5).

Table 1: Schedule sampling. The values in the parentheses is k.

type	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge	METEOR	CIDEr	SPICE	SPIDEr
baseline	0.584	0.394	0.260	0.168	0.402	0.191	0.462	0.137	0.299
linear(0.005)	0.608	0.419	0.280	0.181	0.411	0.195	0.495	0.144	0.319
linear(0.010)	0.614	0.419	0.278	0.181	0.420	0.197	0.496	0.141	0.319
linear(1/90)	0.616	0.423	0.279	0.182	0.418	**0.201**	0.511	**0.147**	0.329
linear(0.015)	0.612	0.418	0.277	0.177	0.416	0.191	0.497	0.138	0.317
linear(0.020)	0.622	0.423	0.276	0.177	0.422	0.197	0.496	0.143	0.320
exp(0.95)	0.615	0.407	0.258	0.153	0.410	0.181	0.450	0.128	0.289
exp(0.96)	0.609	0.410	0.264	0.161	0.413	0.185	0.468	0.134	0.301
exp(0.97)	0.618	0.410	0.265	0.169	0.418	0.188	0.470	0.135	0.303
exp(0.98)	0.612	0.415	0.271	0.171	0.414	0.191	0.468	0.140	0.304
exp(0.99)	0.617	0.421	0.276	0.177	0.417	0.194	0.481	0.143	0.312
sigmoid(9)	0.630	0.412	0.262	0.160	0.413	0.182	0.454	0.130	0.292
sigmoid(11)	0.608	0.410	0.267	0.170	0.419	0.196	0.483	0.139	0.311
sigmoid(13)	0.614	0.422	0.281	0.185	0.418	0.199	0.490	0.144	0.317
sigmoid(15)	0.613	0.420	0.278	0.178	0.417	0.200	0.494	0.145	0.320
sinusoidal(5)	**0.636**	**0.444**	**0.299**	**0.198**	**0.431**	0.198	**0.526**	0.143	**0.334**
sinusoidal(9)	0.619	0.429	0.288	0.190	0.418	0.197	0.515	0.145	0.330
sinusoidal(15)	0.615	0.423	0.280	0.178	0.416	0.195	0.495	0.140	0.317
sinusoidal(20)	0.623	0.432	0.289	0.188	0.422	0.199	0.515	0.145	0.330

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - 14th European Conference, ECCV 2016, Proceedings*, volume Part V of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 382–398, United States, 2016. Springer, Springer Nature. European Conference on Computer Vision (14th : 2016) ; Conference date: 11-10-2016 Through 14-10-2016.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, 2015.

[3] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019.

[4] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[5] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[6] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[8] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.