

# Human Consensus-Oriented Image Captioning

Ziwei Wang, Zi Huang and Yadan Luo

School of Information Technology and Electrical Engineering  
The University of Queensland, Australia

ziwei.wang@uq.edu.au, huang@itee.uq.edu.au, lyadanluo@gmail.com

## Abstract

Image captioning aims to describe an image with a concise, accurate, and interesting sentence. To build such an automatic neural captioner, the traditional models align the generated words with a number of human-annotated sentences to mimic human-like captions. However, the crowd-sourced annotations inevitably come with data quality issues such as grammatical errors, wrong identification of visual objects and sub-optimal sentence focuses. During the model training, existing methods treat all the annotations equally regardless of the data quality. In this work, we explicitly engage human consensus to measure the quality of ground truth captions in advance, and directly encourage the model to learn high quality captions with high priority. Therefore, the proposed consensus-oriented method can accelerate the training process and achieve superior performance with only supervised objective without time-consuming reinforcement learning. The novel consensus loss can be implemented into most of the existing state-of-the-art methods, boosting the BLEU-4 performance by maximum relative 12.47% comparing to the conventional cross-entropy loss. Extensive experiments are conducted on MS-COCO Image Captioning dataset demonstrating the proposed human consensus-oriented training method can significantly improve the training efficiency and model effectiveness.

## 1 Introduction

Visual content understanding has become an emerging research topic in the multimedia research areas recently. In the conventional multimedia retrieval systems, the image or video items can be retrieved by keywords, image snippets or video clips, whilst the textual descriptions can be searched by visual content. Furthermore, in the modern multimedia intelligent system, the underlying semantics in the vision and language are demanding elements for human to efficiently manage big volumes of images and videos. Image captioning system automatically generates descriptions for visual contents, which can benefit applications such as multimedia re-

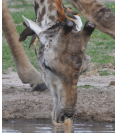

Image	Annotations	CIDEr	BLEU-4
	A giraffe puts its nose into a pool of water	61.89	0.00
	A close up of a giraffe bending down drinking water	<b>141.30</b>	0.01
	A giraffe sticking its face into a watering hole	<b>50.81</b>	<b>0.00</b>
	There is a giraffe that is drinking water	99.50	0.00
	A very cute giraffe bending down to drink some water	96.57	0.00
	An old army truck sitting by a yellow fire hydrant	<b>229.78</b>	38.26
	An old pick up truck parked next to a fire hydrant	129.20	0.00
	An old truck sitting next to a yellow fire hydrant	225.69	45.18
	A photograph of an outside with numerous things in the scene	<b>4.33</b>	<b>0.00</b>
	A yellow fire hydrant next to an army truck	213.40	37.17

Figure 1: Human-annotated captions inevitably come with label quality issues, but the existing methods treat them equally during training. Our proposed consensus-oriented captioning model explicitly diminishes the inherent label noise. Up: The sentence in red box only bluntly describes the objects without revealing the interactions such as “drinking”, “bending down”. Bottom: Then sentence in red box is the longest caption, but the quality is the poorest due to the lack of information. BLEU-4 metric fails to distinguish the quality in these cases.

trieval, data management and recommender system [Li *et al.*, 2020], etc. Notably, the challenging image captioning task requires the model to accurately identify and interpret visual features into high-level natural language descriptions. The current approaches for image captioning follow the encoder-decoder architecture. In particular, the convolutional neural network (CNN) encoder projects the image into a feature vector conserving the salient elements, and the decoder language model generates natural language descriptions by maximising the posterior probability of word predictions given the image representations.

In the standard MSCOCO Image Captioning dataset, each image comes with five annotations shown in Figure 1. The captions are annotated by different online annotators. Although they are all human-annotated captions, the quality is apparently different. The captions may be correct, but they sometimes come with poor grammar and lack of information. For example, in Figure 1, a good example like the second sentence gives all the details about the giraffe and its interaction of drinking water, but the bad annotation such as the third caption only describes the pure motion of bending down. In addition, instead of describing the items from the image, the

annotators may write some “one-caption-fit-all” captions. In the bottom example of Figure 1, the fourth sentence is the longest, but it does not include any useful information from the given visual content. During the training, the existing models are encouraged to generate captions via supervised objectives. They treat all the annotations of the same image equally during training without considering the quality of different labels.

The successive work mainly focuses on improving the performance with fine-grained visual representations. There are several ways to improve visual encodings, for example, the objects from the images are detected to form a detailed description [Herdade *et al.*, 2019], the visual attributes of items are classified to accurately describe objects [Yao *et al.*, 2017], and the visual attention mechanisms are equipped to focus on salient areas while generating captions [Xu *et al.*, 2015; Anderson *et al.*, 2018]. However, despite the benefits from the visual information, the cross-entropy objectives encourage the model to fully rely on the ground-truth annotations, lacking the self-critical evaluation mechanism to adjust the gradient.

Built on the encoder-decoder baselines, the policy gradient reinforcement learning captioning models [Rennie *et al.*, 2017; Zha *et al.*, 2019] are further proposed to minimise the gap between the cross-entropy loss and the evaluation metrics. The captioning task is formulated as a sequential decision-making problem, in which the language policy are directly optimised based on rewards from the caption evaluator. However, in practice, the reinforcement learning cannot start from scratch, due to the fact that the action searching space is tremendous with random initialisation. The normal practice is to “warm-up” the policy network with supervised training, followed by very slow fine-tuning process with reinforcement learning loss. This process is time-consuming, with turbulent learning curve and the performance is over-sensitive to network initialisation.

To alleviate the intrinsic annotation imperfect and model fine-tuning inefficient deficiencies, in this work, we introduce a Human Consensus-Oriented (HCO) objective to improve model performance with faster training. The proposed consensus loss can automatically assess human annotation quality in advance, aiming to encourage the model to learn more accurate and informative training samples in priority. To evaluate the effectiveness of the proposed training objective, we perform comprehensive experiments on both basic and advance models, and demonstrate prominent improvements on different metrics.

The contributions in this paper are three-fold:

1. To the best of our knowledge, it is the first work to explicitly identify and tackle the annotation quality issues in image captioning, which allows the model to efficiently learn the higher quality annotations in priority.
2. The proposed consensus loss is agnostic to the base model, so it can be easily implemented into most of the existing methods to improve the performance with supervised training.
3. Quantitative and qualitative experiments are conducted on the challenging COCO Captioning dataset. The re-

sults show that in the supervised learning setting, our model outperforms all the cross-entropy trained baseline models. In the reinforcement learning fine-tuning, the method still achieves competitive performance comparing the state-of-the-arts.

The rest of the paper is organised as follows: Section 2 reviews visual captioning methods, Section 3 introduces the consensus loss and the framework, Section 4 describes experiments, and Section 5 summarises the proposed method.

## 2 Related Work

### 2.1 Visual Captioning

Recently, visual captioning has been widely studied in the computer vision and natural language processing fields. The existing deep captioning models follow the encoder-decoder framework, aiming to map the visual content into a full sentence [Karpathy and Fei-Fei, 2017] or a paragraph [Krause *et al.*, 2017; Wang *et al.*, 2018; Luo *et al.*, 2019], given the image or video [Gao *et al.*, 2019; Yang *et al.*, 2018; Song *et al.*, 2018] representations. Benefiting from the feature extraction ability of convolutional neural networks, the early captioning work focuses on preserving effective image representations to generate high quality descriptions. In [Vinyals *et al.*, 2015], Vinyals *et al.* proposed a CNN-LSTM architecture to recurrently predict the caption word-by-word to form a full sentence based on the CNN-encoded image features.

### Attention Model

The successive models focus on improving the visual representations by different attention mechanism designs. The attention mechanism can be categorised into three types: the visual attention [Xu *et al.*, 2015], language attention [You *et al.*, 2016], and hybrid attention [Lu *et al.*, 2017]. The visual attention [Xu *et al.*, 2015; Anderson *et al.*, 2018] aligns the salient image areas while generating word sequence to improve the relevance of the generated caption. The language attention [You *et al.*, 2016] detects semantics of the given image, and then the decoder refer to these semantic attributes when generating the captioning. The hybrid methods [Lu *et al.*, 2017] combines visual and textual attention to simultaneously consider both types of information. The more recent attention-based models take a step further to entirely utilise the Transformer attention [Herdade *et al.*, 2019; Wang *et al.*, 2020] to encode visual features and decode word sequences.

### Reinforcement Learning

Another line of work formulate image captioning as a reinforcement learning problem [Rennie *et al.*, 2017; Zha *et al.*, 2019; Luo *et al.*, 2019]. In the reinforcement learning setting, the action is predicting next word, the state is the visual features, current word embeddings and context vectors, and the reward is the evaluation score of a group of sampled captions. Most of the existing methods need to warm-up the network with supervised training, then utilise the REINFORCE [Williams, 1992] algorithm to optimise the gradient for model fine-tuning. The SCST model [Rennie *et al.*, 2017]

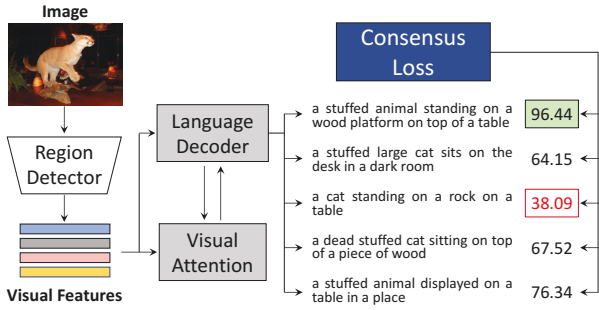


Figure 2: The overview of the proposed human consensus-oriented captioner. Annotations with higher quality are encouraged to be learned in priority.

further introduces a baseline subtracted from the reward to reduce the variance caused by Monte-Carlo sampling. In the CAVP model [Zha *et al.*, 2019], the model improves the visual policy with previous context, and optimises the network with the Actor-Critic policy gradient method.

### 2.2 Dataset Bias

The deep intelligence systems rely on large volume of training data, so the bias in data could be magnified during gradient descent. It is commonly existed in varies tasks such as object detection, word embedding, and image captioning. In particular, in the object detection, the RetinaNet [Lin *et al.*, 2018] tackles the class imbalance problem by down-weighting the loss of already well-classified instances, therefore guiding the model to learn more on the hard examples. The word embedding model [Bolukbasi *et al.*, 2016] identifies and equalises the gender-related and gender-neutral words, while maintaining good clustering performance. The image captioning models [Hendricks *et al.*, 2018] also focus on gender-specific word generates, and they increase the penalty for the wrongly generated gender words while model training.

## 3 The Proposed Approach

In this section, we introduce the proposed Human Consensus-Oriented (HCO) image captioning model. As shown in the Figure 2, during the encoding phase, the object regions in the input image are firstly detected by Faster-RCNN. Then, in the decoding phase, the language model generates word sequences based on the visual representation. During the word prediction, the visual attention is engaged to shift the sentence focus on the salient object regions.

### 3.1 Problem Formulation

We denote the image captioning task formally in this section. The original RGB image is denoted as the input  $I$ . The final objective of the image captioning model is to generate a description  $S = \{S_1, \dots, S_N\}$  given  $I$ , where  $N$  is the length of caption.

### 3.2 Consensus Loss

The Consensus Loss (CL) is designed to address the annotation quality issue, in which the annotated captions have different levels of quality. In the training dataset, each image has

several captions from different annotators as shown in Figure 1. The existing methods force the model to learn these captions equally without considering the quality of themselves. The proposed consensus loss explicitly puts different learning weights for different quality of training examples.

### Ground Truth Consensus Scores

We denote the reference captions of the given image  $I$  as  $\mathbf{R}^I = \{\mathbf{R}_1^I, \dots, \mathbf{R}_M^I\}$ , where  $M$  is the number of ground-truth annotations of image  $I$ . Before the model training, we split  $\mathbf{R}^I$  into  $M$ -folds. For each caption  $\mathbf{R}_i^I$ , we measure the consensus score  $C_{\text{score}}(\mathbf{R}_i^I)$  based on the common interest among all the other  $M - 1$  human annotators. These  $M - 1$  references can be denoted as  $\mathbf{Q}_i^I$  for simplicity.

The choice of consensus score measurement is not limited to the existing metrics, however, in practice, we adopt CIDEr [Vedantam *et al.*, 2015] to quantify the label quality. CIDEr calculates the cosine similarity between the candidate sentence and a set of reference sentences. In particular, the sentence is represented by a vector of Term Frequency Inverse Document Frequency (TF-IDF) weightings for all the  $n$ -gram phrase, where  $n = 1, 2, 3, 4$  in practice. Finally, all the  $n$ -gram cosine similarity are averaged to form a single score for each candidate sentence.

The  $C_{\text{score}}(\mathbf{R}_i^I)$  is computed as follows:

$$\text{CIDEr}_n(\mathbf{R}_i^I, \mathbf{Q}_i^I) = \frac{1}{M - 1} \sum_j \frac{r^n(\mathbf{R}_i^I) \cdot r^n(\mathbf{Q}_{ij}^I)}{\|r^n(\mathbf{R}_i^I)\| \|r^n(\mathbf{Q}_{ij}^I)\|} \quad (1)$$

$$C_{\text{score}}(\mathbf{R}_i^I, \mathbf{Q}_i^I) = \sum_{n=1}^4 \text{CIDEr}_n(\mathbf{R}_i^I, \mathbf{Q}_i^I), \quad (2)$$

where  $r^n$  is the TF-IDF weightings of all the  $n$ -gram phrases for the corresponding sentence.

### Cross-Entropy with Consensus Loss

The standard cross-entropy (CE) loss measures the classification model performance based on the probability outputs from the model. The loss of the entire sentence is measured as the mean of all the words' CE loss. The standard CE loss simply treats all the annotations as golden standards, and gives different quality captions with the same weights.

The consensus loss adjusts the sentence-level loss, which increases the weights of high quality examples, but reduces the loss of the low quality ones. Therefore, higher penalty will be given to model if the ‘‘good’’ captions are not learned well. Meanwhile, the poorly written captions will be assigned with lower loss, so the model will not learn too much about bad examples. This follows the common practice of human learning behaviour: We learn well-written articles in the school first, and later we have the ability to identify poorly-written negative examples.

We write the simple Consensus Loss (CL) as:

$$\text{CL}(I, S) = -C_{\text{score}}(\mathbf{R}_i^I) \frac{1}{N} \sum_{t=1}^N \log p_t(S_t), \quad (3)$$

where  $N$  is the length of caption,  $\mathbf{R}_i^I$  is the  $i$ -th annotation for image  $I$ ,  $p_t \in [0, 1]$  is the likelihood for the correct word  $S_t$ .

Intuitively, the sentence with poor quality will have low loss contribution, whilst the high quality examples are learned in priority benefited from the high loss contribution. The loss of the perfect learned sentences will have zero CL loss, so it will not be contributing to model training any more despite its consensus score.

### Variants of Consensus Loss

The extract composition and acquisition of consensus loss are not fixed, we introduce some alternative designs of the consensus loss. The different weights of CL loss and consensus evaluation metrics are compared in Section 4.5 and Section 4.6, respectively.

**Balancing Factor.** The proportion of consensus loss can be further adjusted empirically to balance the standard cross-entropy and consensus loss. We use the Consensus Loss with balancing factor  $\alpha$  in practice:

$$CL(I, S) = -(1 + \alpha C_{score}(\mathbf{R}_i^I)) \cdot \frac{1}{N} \sum_{t=1}^N \log p_t(S_t) \quad (4)$$

**Label Quality Measurements.** The calculation of consensus score can be reshaped to other automatic evaluation metrics or expert annotations. The commonly adopted metrics automatically measure the similarity of the candidate caption with reference annotations, such as BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], CIDEr [Vedantam *et al.*, 2015], etc.

### 3.3 Human Consensus-Oriented Captioner

The consensus loss can be implemented in most of the image captioning models trained by cross entropy loss. In this section, we briefly introduce two consensus-based models built on CNN-LSTM and Transformer framework.

#### CNN-LSTM

We implement consensus loss on the state-of-the-art Top-down attention-based model [Anderson *et al.*, 2018]. The visual features are extracted from Faster-RCNN, and the LSTM language decoder is equipped with visual attention to adaptively shift sentence focus.

#### Transformer

Similar to CNN-LSTM model, the visual features are extracted from Faster-RCNN, however, the Transformer model [Vaswani *et al.*, 2017] utilises a self-attention mechanism to generate sentence without using recurrent units. During decoding phase, the visual region features are attended to calculate the relevance scores and output the context vectors.

## 4 Experiments

### 4.1 Experimental Settings

#### Dataset

We evaluate the proposed HCO model on the MS-COCO [Lin *et al.*, 2014] image captioning dataset following the ‘‘Karpathy’’ [Karpathy and Fei-Fei, 2017] split. The Train, Val, Test splits contain 113 287, 5 000, 5 000 images, respectively.

### Evaluation Metrics

The performance is evaluated using automatic language evaluation metrics: BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], Rouge-L [Lin, 2004], and CIDEr [Vedantam *et al.*, 2015]. BLEU- $n$  indicates the  $n$ -gram precision of the candidate caption, METEOR considers both precision and recall, and CIDEr measures the  $n$ -gram similarity with considering TF-IDF weights.

### 4.2 Baselines

We compare the proposed HCO with a number of state-of-the-art methods (see Table 1) including: NIC [Vinyals *et al.*, 2015], Adaptive [Lu *et al.*, 2017], Att2all and Att2in [Rennie *et al.*, 2017], Topdown [Anderson *et al.*, 2018], Standard Transformer and Object Transformer [Herdade *et al.*, 2019], CAVP [Zha *et al.*, 2019].

In the following comparison, **CL-** indicates reshaping the original Cross-Entropy Loss to the proposed Consensus Loss. **RL-** denotes the model is optimised with Reinforcement Learning (Self-critical model from [Rennie *et al.*, 2017]). All the models with region visual features use Faster-RCNN with ResNet-101 [Anderson *et al.*, 2018]. The specific variations will be discussed in the following sections.

### 4.3 Implementation Details

The RGB image is encoded with ResNet-101 convolutional neural network, and the regions are detected via Faster-RCNN. The model is optimised using Adam with learning rate of  $5e-4$ . Different settings are implemented in the CNN-LSTM and Transformer architectures. The details will be released in the Github repository.

**CNN-LSTM.** In the LSTM language decoder, the hidden state is empirically set as 1024, with 1-layer. In the attention module, the encoding size is 1024.

**Transformer.** In the Transformer attention framework, the embedding dimension is 512, the positional encoding size is 2048, and the number of attention layer is 6.

### 4.4 Comparison with State-of-The-Art

#### Quantitative Analysis

We compare our model with state-of-the-art methods on MSCOCO dataset. All the visual features are using Faster-RCNN region features for fair comparison, so some of the reported results are slightly higher than the original papers. As

Method	Base	Attention	Feature
NIC	CNN-LSTM	No	Global
Adaptive	CNN-LSTM	Yes	Region
Att2all	CNN-LSTM	Yes	Region
Att2in	CNN-LSTM	Yes	Region
Topdown	CNN-LSTM	Yes	Region
CAVP	CNN-LSTM	Yes	Region
Transformer	Transformer	Yes	Region
Obj. Trans.	Transformer	Yes	Region, Obj. Relation

Table 1: Comparison methods summary.

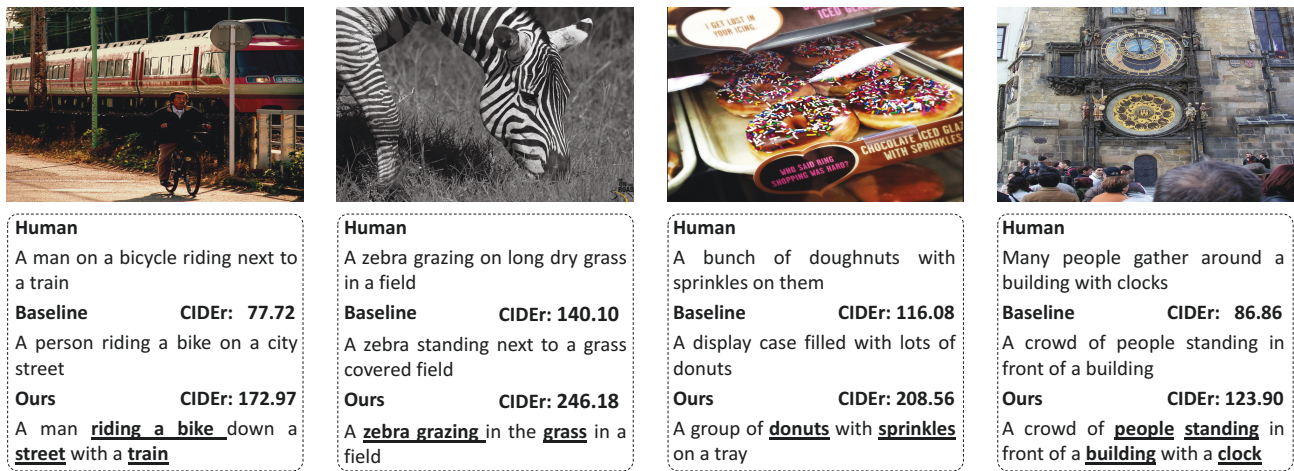


Figure 3: Case studies of HCO, Topdown baseline, and human annotations in different scenes.

shown in the Table 2, all the compared models are trained on the standard cross-entropy (CE) loss without reinforcement learning. Our CL-trained model CL-Topdown improves the CE-trained Topdown baseline with 3.2% relatively improvement on CIDEr. It also surpasses the more recent state-of-the-art model Object Transformer model, which is very significant because the Object Transformer method utilises more powerful Transformer model and uses auxiliary region box relationship features. Another comparison demonstrated in the Table 3 compare the models after the reinforcement learning fine-tuning process. The RL-optimised CL-RL-Topdown further improves the CIDEr by 5.27, and also performs better than the original RL-Topdown. This shows that the CL loss can also provides self-critical RL training with a better starting point.

### Qualitative Analysis

We present case studies in Figure 3 to intuitively understand the HCO model performance. The proposed model is able to generate sentences with accurate observation of salient objects from the images in most of the cases. Besides, comparing to the baseline method, it is able to describe more interesting elements. For example, in the first picture, rather than only describing the rider on the street, our model finds the interesting point “train”, and gains higher CIDEr score

Model	B-4	M	Rouge	CIDEr
NIC	30.34	25.05	53.58	96.29
Adaptive	30.88	25.40	53.82	98.26
Att2in	31.83	25.73	54.50	102.29
Att2all	33.25	26.25	55.19	105.60
Topdown	36.20	27.00	56.40	113.50
Transformer	34.96	27.58	55.82	111.89
Obj. Trans.	35.49	27.98	56.58	115.37
<b>CL-Topdown</b>	<b>37.08</b>	<b>27.85</b>	<b>57.22</b>	<b>117.10</b>
<b>CL-Transformer</b>	<b>35.65</b>	<b>27.68</b>	<b>56.42</b>	<b>114.75</b>

Table 2: Performance comparison on MSCOCO Karpathy test split w/o reinforcement learning

since it is the human annotators’ common interest. Similar phenomenons also appear in the third and fourth pictures, in which the interesting elements of “sprinkles” and “clock” are accurately identified. Moreover, in the second picture, it not wrong that the zebra is “standing”, but the motion of “grazing” is more accurate. In addition, the grammar and language usages of the generated captions are more appropriate.

## 4.5 Ablation Study

### Effectiveness of Consensus Loss

In this section, a number of state-of-the-art methods are implemented with consensus loss to compare the effectiveness of the consensus loss (Table 4). All the CNN-LSTM based models (see Table 1) are implemented with Faster-RCNN visual features, and the hidden size of 1-layer LSTM is 512. The Transformer model parameters remain the same.

From the comparison, we can clearly observe that the consensus loss works in all the compared methods without auxiliary visual information. In particular, the CL loss boosts BLEU-4 score of the Att2in model by 12.47 % relatively. Similarly, the Adaptive model has a giant relative increment of 12.08 % and 9.75 % for BLEU-4 and CIDEr, respectively. Notably, when the standard Transformer model utilises CL loss, the performance nearly achieves the heavy Object Transformer model.

Model	B-4	M	Rouge	CIDEr
RL-Att2all	34.20	26.70	55.70	114.00
RL-Topdown	36.30	27.70	56.90	120.10
RL-Obj. Trans.	38.60	28.70	58.40	128.30
RL-CAVP	38.60	28.30	58.50	126.30
<b>CL-RL-Topdown</b>	<b>36.56</b>	<b>27.56</b>	<b>57.31</b>	<b>122.37</b>
<b>CL-RL-Transformer</b>	<b>37.13</b>	<b>28.23</b>	<b>57.65</b>	<b>126.13</b>

Table 3: Performance comparison on MSCOCO Karpathy test split w/ reinforcement learning

Model	Cross-Entropy				Consensus Loss				Improvement	
	B-4	M	Rouge-L	CIDEr	B-4	M	Rouge-L	CIDEr	B-4	CIDEr
NIC	30.34	25.05	53.58	96.29	32.95	25.54	54.32	101.84	8.60 %	5.76 %
Adaptive	30.88	25.40	53.82	98.26	34.61	26.45	55.58	107.84	12.08 %	<b>9.75 %</b>
Att2in	31.83	25.73	54.50	102.29	35.80	26.87	56.33	111.47	<b>12.47 %</b>	8.97 %
Att2all	33.25	26.25	55.19	105.60	36.36	27.05	56.64	112.97	9.35 %	6.98 %
Topdown	33.14	26.45	55.37	106.14	36.52	27.42	56.90	114.22	10.20 %	7.61 %
Trans.	34.96	27.58	55.82	111.89	35.65	27.68	56.42	114.75	1.97 %	2.56 %
Obj. Trans.	35.49	27.98	56.58	115.37	36.25	28.10	56.71	117.73	2.14 %	2.05 %

Table 4: Effectiveness of Consensus Loss. The models remain the identical structure, but the loss function is replaced with consensus loss.

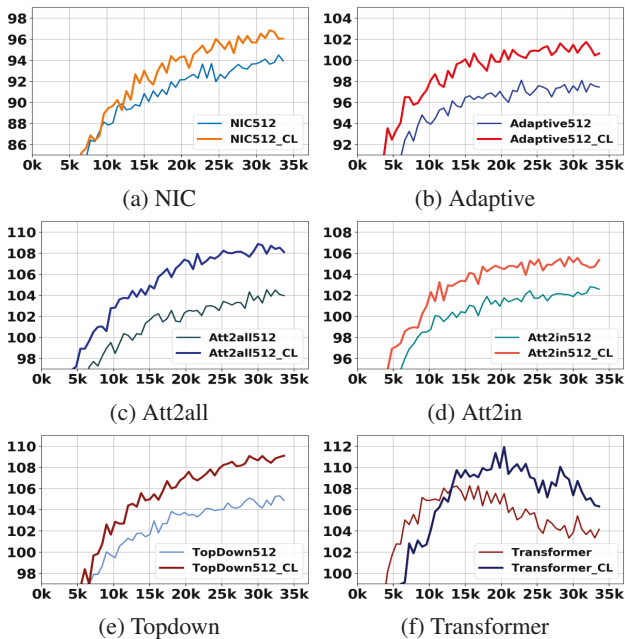


Figure 4: Training comparison of consensus loss. X-axis: Training steps; Y-axis: CIDEr score in Val split.

### Efficiency of Model Training

In this section, we study the training efficiency of CL loss comparing to the standard cross-entropy loss illustrated in the Figure 4. In the graph, the X-axis is the training steps, in which each step contains a mini-batch of 100 examples, and the Y-axis is the CIDEr score of the validation set with greedy search. In most of the cases, the model can achieve the original performance in only third to half of the training steps without extra training cost.

### Model Structure Comparison

In section 3.2, we introduce the balancing factor  $\alpha$  to regulate the contribution of consensus loss. As shown in the Table 5, when the ratio of cross-entropy and consensus loss is 1 : 2, the model performs the best in METEOR and CIDEr metrics, and other metrics are competitive to other settings.

### 4.6 Label Quality Measurements Comparison

As described in the Section 3.2, the training performance of different consensus loss measurements are compared. In the

Model	B-4	M	Rouge	CIDEr
$\alpha = 0.5$	36.55	27.76	56.89	114.60
$\alpha = 1$	<b>37.26</b>	28.00	57.20	116.42
$\alpha = 2$	37.20	<b>28.03</b>	57.10	<b>117.51</b>
$\alpha = 5$	36.66	27.86	57.02	115.67
$\alpha = 10$	37.12	27.85	57.02	116.84
Pure CL	37.08	27.85	<b>57.22</b>	117.10

Table 5: Balancing factor comparison. Performance comparison on MSCOCO Karpathy test split. The backbone model is Topdown.

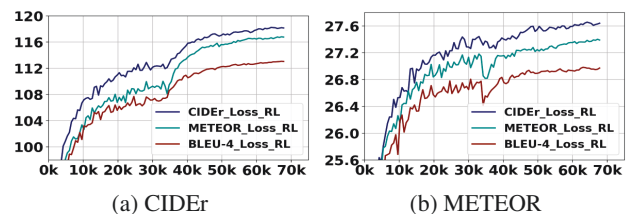


Figure 5: Variants of consensus measurement. X-axis: Training steps; Y-axis: Evaluation scores in Val split.

Figure 5, the 5a is the CIDEr scores, the 5b is the METEOR scores. We train the model with different CL in the first 35k steps, and then optimise the model with self-critical reinforcement learning [Rennie *et al.*, 2017]. In the figures, the different lines indicate different CL choices. We can find that the CIDEr consensus loss works well in both metrics, and it also attains a smoother training curve comparing to other metrics.

## 5 Conclusion and Future Work

In this work, we propose a human consensus-oriented model for image captioning. Towards retaining common interests among the references, the proposed model explicitly leverages the consensus scores to encourage captioner to learn high quality examples in priority. The model boosts the caption generation performance without overwhelming training time effectively, and generates the concise, accurate and interesting examples efficiently. The experiments demonstrate the effectiveness of the proposed consensus-oriented captioner.

## Acknowledgements

This work is partially supported by ARC DP190102353 and ARC DP170103954.

## References

- [Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.
- [Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72, 2005.
- [Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, 2016.
- [Gao *et al.*, 2019] L. Gao, X. Li, J. Song, and H. T. Shen. Hierarchical lstms with adaptive attention for visual captioning. *TPAMI*, 2019.
- [Hendricks *et al.*, 2018] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pages 793–811, 2018.
- [Herdade *et al.*, 2019] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.
- [Karpathy and Fei-Fei, 2017] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *TPAMI*, pages 664–676, 2017.
- [Krause *et al.*, 2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, pages 3337–3345, 2017.
- [Li *et al.*, 2020] Yang Li, Yadan Luo, and Zi Huang. Graph-based relation-aware representation learning for clothing matching. In *ADC*, pages 189–197, 2020.
- [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.
- [Lin *et al.*, 2018] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *TPAMI*, 2018.
- [Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. ACL, 2004.
- [Lu *et al.*, 2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
- [Luo *et al.*, 2019] Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Jingjing Li, and Yang Yang. Curiosity-driven reinforcement learning for diverse visual paragraph generation. *ACM MM*, pages 2341–2350, 2019.
- [Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.
- [Rennie *et al.*, 2017] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195, 2017.
- [Song *et al.*, 2018] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. From deterministic to generative: Multimodal stochastic rnns for video captioning. *TNNLS*, pages 3047–3058, 2018.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017.
- [Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
- [Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.
- [Wang *et al.*, 2018] Ziwei Wang, Yadan Luo, Yang Li, Zi Huang, and Hongzhi Yin. Look deeper see richer: Depth-aware image paragraph captioning. In *ACM MM*, pages 672–680, 2018.
- [Wang *et al.*, 2020] Ziwei Wang, Zi Huang, and Yadan Luo. PAIC: Parallelised attentive image captioning. In *ADC*, pages 16–28, 2020.
- [Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, pages 229–256, 1992.
- [Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.
- [Yang *et al.*, 2018] Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, H. T. Shen, and Yanli Ji. Video captioning by adversarial lstm. *TIP*, pages 5600–5611, 2018.
- [Yao *et al.*, 2017] Ting Yao, Yingwei Pan, Yehao Li, Zhao-fan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, pages 4904–4912, 2017.
- [You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.
- [Zha *et al.*, 2019] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *TPAMI*, 2019.