1.The Traffic Scene Understanding and Prediction Based on Image Captioning

Authors: WEI LI, ZHAOWEI QU, HAIYU SONG, PENGJIE WANG, AND BO XUE

Dataset: Flickr30k

Models used:NIC(VGG16,LSTM)

-NIC(Neural Image Caption) model combining VGG16 for feature extraction and LSTM for sequential caption generation. This model achieved a BLEU score of 64.7. They created a dataset due to the lack of an open one and compared their method with others on Flickr30K. Their approach showed similar performance, with further improvement by fine-tuning their model on their dataset.

TABLE 2. Performance of our method on flickr30K dataset in terms of BLEU scores.

methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC[41]	66.3	42.3	27.7	18.3
Att-NIC[43]	**66.9**	43.9	29.6	19.9
ours	64.7	**46.4**	**31.2**	**22.4**

2.Automatic Image Caption Generation Using Deep Learning

Authors: Akash Verma, Arun Kumar Yadav, Mohit Kumar and Divakar Yadav,

Dataset:MSCOCO,Flickr8k

Evaluation metrics:BLEU,GLEU,ROUGE L and METEOR

Models used:VGG16 Hybrid Places1365,LSTM

- VGG16 Hybrid Places 1365 as an encoder and LSTM as a decoder to create accurate captions for images. After training on labelled Flickr8k and MSCOCO Captions datasets, the model showed promising results, achieving BLEU-1 scores of 0.6666 and 0.7350, METEOR scores of 0.5060 and 0.4768, and GLEU scores of 0.2469 and 0.2798, respectively. These scores surpass those of existing methods. Moreover, the model's effectiveness is confirmed through caption generation on live sample images, reinforcing its performance.

Table 2 BLEU scores of proposed models on Flickr8k dataset. Emphasis indicates result of best model

S.No.	Encoder	Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	Inception V3	LSTM	0.6400	0.4131	0.2809	0.1571
2	Inception V3	B-LSTM	0.6121	0.3733	0.2465	0.3444
3	VGG16	LSTM	0.6400	0.3771	0.2985	0.1501
4	VGG16	B-LSTM	0.6274	0.3873	0.2670	0.1501
5	VGG16 Places365	LSTM	0.6382	0.4088	0.3040	0.2028
6	VGG16 Places365	B-LSTM	0.5956	0.2384	0.1484	0.2345
7	**VGG16 Hybrid Places1365**	LSTM	**0.6666**	**0.4340**	**0.3893**	**0.2878**
8	VGG16 Hybrid Places1365	B-LSTM	0.6441	0.3136	0.1760	0.2682
9	ResNet50	LSTM	0.5543	0.2863	0.1628	0.0877
10	VGG19	LSTM	0.5912	0.3321	0.2527	0.1732

3.Neural Image Caption Generation with Weighted Training and Reference

Authors: Guiguang Ding, Minghai Chen, Sicheng Zhao, Hui Chen,Jungong Han,Qiang Liu

Dataset:MSCOCO,Flickr30k

Models used: ResNet-101,R-LSTM

-It uses the encoder-decoder framework, employing ResNet-101 for image feature extraction, followed by R-LSTM for caption generation. Additionally, comparing with weighted training alone confirms the effectiveness of reference-based generation in R-LSTM. Tests on MS COCO and Flickr30k datasets show R-LSTM achieving a notable 10.37% increase in CIDEr on MS COCO.

Table 3 Performance (%) of the proposed model compared with several state-of-the-art methods on Flickr30k dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Google NIC [59]	66.3	42.3	27.7	18.3	-	-	-
m-RNN [44]	60.0	41.0	28.0	19.0	-	-	-
LRCN [12]	58.7	39.1	25.1	16.5	-	-	-
Toronto [61]	66.9	43.9	29.6	19.9	18.5	-	-
ATT [65]	64.7	46.0	32.4	**23.0**	18.9	-	-
SCA-CNN [3]	66.2	46.8	32.5	22.3	**19.5**	-	-
R-LSTM (ours)	**67.7**	**48.0**	**32.6**	22.1	**19.5**	**45.7**	**45.0**

Table 4 Performance (%) of the proposed model compared with several state-of-the-art methods on MS COCO dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Google NIC [59]	66.6	45.1	30.4	20.3	-	-	-
m-RNN [44]	67.0	49.0	35.0	25.0	-	-	-
LRCN [12]	66.9	48.9	34.9	24.9	-	-	-
Toronto [61]	71.8	50.4	35.7	25.0	23.0	-	-
ATT [65]	70.9	53.7	40.2	30.4	24.3	-	-
USC [21]	69.7	51.9	38.1	28.2	23.5	50.9	83.8
SCA-CNN [3]	71.9	54.8	41.1	31.1	25.0	-	-
GLA-BEAM3 [34]	72.5	55.6	41.7	31.2	24.9	53.3	96.4
R-LSTM (ours)	**76.5**	**60.3**	**45.8**	**34.4**	**26.4**	**55.7**	**106.4**

4. Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction

Authors: Reshmi Sasibhooshan , Suresh Kumaraswamy and Santhoshkumar Sasidharan

Dataset:MSCOCO,Flickr30k, Flickr8k

Models used:WCNN,LSTM

-It uses an encoder-decoder framework along with techniques like Visual Attention Prediction Network (VAPN) and Contextual Spatial Relation Extractor (CSE) network. The model is trained on three datasets: Flickr8K, Flickr30K, and MSCOCO. It achieves a high CIDEr score of 124.2 on the MSCOCO dataset, showing its effectiveness in creating detailed captions for images.

Table 7 Results of ablation study conducted on MSCOCO dataset

Configuration	Cross-Entropy loss		Self-Critical loss	
	B@4	CD	B@4	CD
WCNN+atr+LSTM	33.1	109.2	34.4	116.5
WCNN+atr+SA+LSTM	33.9	110.8	35.7	117.9
WCNN+atr+SA+CA+LSTM	35.2	112.7	36.3	119.0
WCNN+atr+CA+SA+LSTM	35.9	113.4	37.1	120.4
WCNN+atr+CA+SA+CSE+LSTM	37.5	116.9	38.2	124.2

Here *atr* atrous convolution, *SA* spatial attention, *CA* channel attention

5.Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users

Authors: Yashwant Dongare, Dr. Bhalchandra M. Hardas, Dr. Rashmita Srinivasan, Dr. Vidula Meshram, Dr. Mithun G. Aush, Dr. Atul Kulkarni

Dataset:MSCOCO,Flickr30k

Model Used: Hybrid CNN+LSTM,CNN,LSTM.

-The dataset includes MS COCO, Flickr8k, Flickr30k, PASCAL 1K, and AI Challenger Dataset. Image Caption Generation with Attention Mechanism is employed, featuring an Encoder with convolutional features and a Decoder with a Long Short-Term Memory (LSTM) network. The results are shown below.

Table 3: Comparative summary of assessment of proposed model MS COCO Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
CNN+LSTM	99.42	99.41	99.63	99.55
CNN	96.01	97.56	97.64	98.72
RNN	95.26	97.74	99.57	98.53
DNN	95.87	97.93	98.93	98.17
DBN	97.29	91.37	97.76	95.54