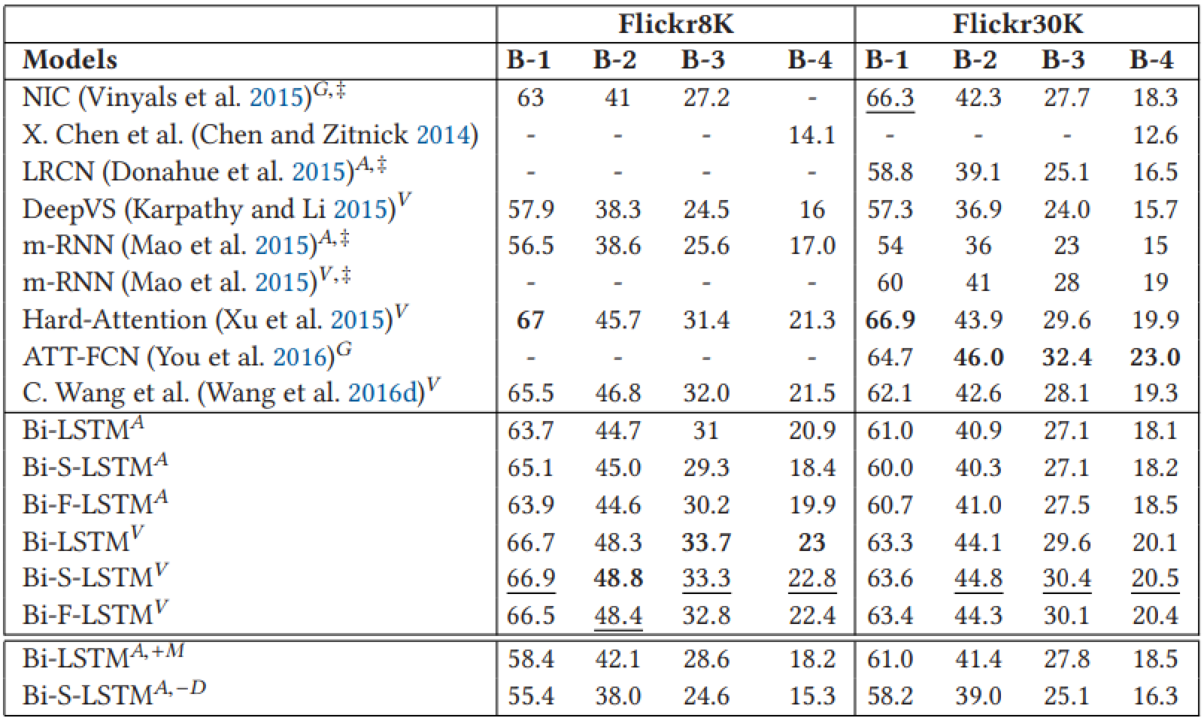
1) Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning

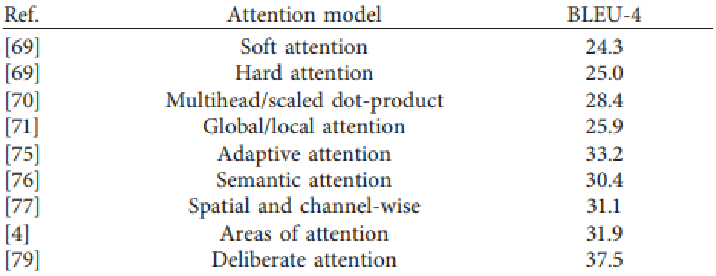
- **Authors: CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL, Hasso Plattner Institute, University of Potsdam**
- Dataset used Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets.
- Done using by combining a **deep convolutional neural network (CNN)** and **two separate LSTM networks**



	Flickr8K				**Flickr30K**			
Models	**B-1**	**B-2**	**B-3**	**B-4**	**B-1**	**B-2**	**B-3**	**B-4**
NIC (Vinyals et al. 2015)^{G,‡}	63	41	27.2	-	66.3	42.3	27.7	18.3
X. Chen et al. (Chen and Zitnick 2014)	-	-	-	14.1	-	-	-	12.6
LRCN (Donahue et al. 2015)^{A,‡}	-	-	-	-	58.8	39.1	25.1	16.5
DeepVS (Karpathy and Li 2015)^{V}	57.9	38.3	24.5	16	57.3	36.9	24.0	15.7
m-RNN (Mao et al. 2015)^{A,‡}	56.5	38.6	25.6	17.0	54	36	23	15
m-RNN (Mao et al. 2015)^{V,‡}	-	-	-	-	60	41	28	19
Hard-Attention (Xu et al. 2015)^{V}	**67**	45.7	31.4	21.3	**66.9**	43.9	29.6	19.9
ATT-FCN (You et al. 2016)^{G}	-	-	-	-	64.7	**46.0**	**32.4**	**23.0**
C. Wang et al. (Wang et al. 2016d)^{V}	65.5	46.8	32.0	21.5	62.1	42.6	28.1	19.3
Bi-LSTM^{A}	63.7	44.7	31	20.9	61.0	40.9	27.1	18.1
Bi-S-LSTM^{A}	65.1	45.0	29.3	18.4	60.0	40.3	27.1	18.2
Bi-F-LSTM^{A}	63.9	44.6	30.2	19.9	60.7	41.0	27.5	18.5
Bi-LSTM^{V}	66.7	48.3	**33.7**	**23**	63.3	44.1	29.6	20.1
Bi-S-LSTM^{V}	66.9	**48.8**	33.3	22.8	63.6	44.8	30.4	20.5
Bi-F-LSTM^{V}	66.5	48.4	32.8	22.4	63.4	44.3	30.1	20.4
Bi-LSTM$^{A,+M}$	58.4	42.1	28.6	18.2	61.0	41.4	27.8	18.5
Bi-S-LSTM$^{A,-D}$	55.4	38.0	24.6	15.3	58.2	39.0	25.1	16.3

2) An Overview of Image Caption Generation Methods

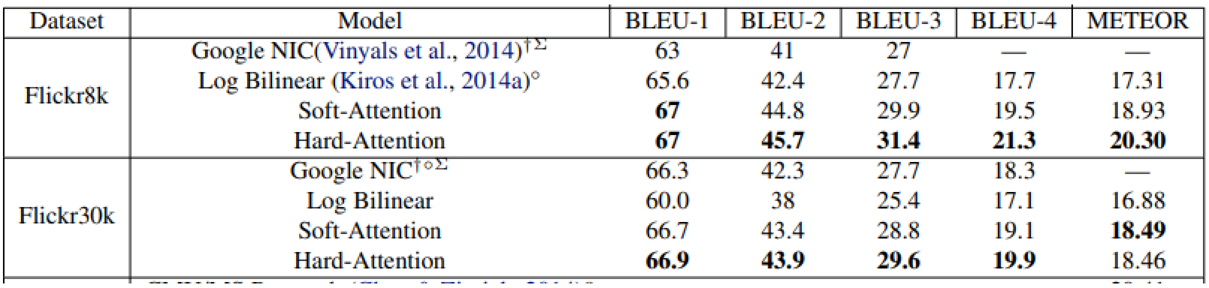
- **Authors: Haoran Wang , Yue Zhang, and Xiaosheng Yu**
- Dataset used MSCOCO, Flickr8k, Flickr30k, PASCAL 1K, AI Challenger Dataset
- Done using Image Caption Generation with **Attention Mechanism**(ENCODER: CONVOLUTIONAL FEATURES),(DECODER: LONG SHORT-TERM MEMORY NETWORK)
- **BLEU Score:**



Ref.	Attention model	BLEU-4
[69]	Soft attention	24.3
[69]	Hard attention	25.0
[70]	Multihead/scaled dot-product	28.4
[71]	Global/local attention	25.9
[75]	Adaptive attention	33.2
[76]	Semantic attention	30.4
[77]	Spatial and channel-wise	31.1
[4]	Areas of attention	31.9
[79]	Deliberate attention	37.5

3)Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

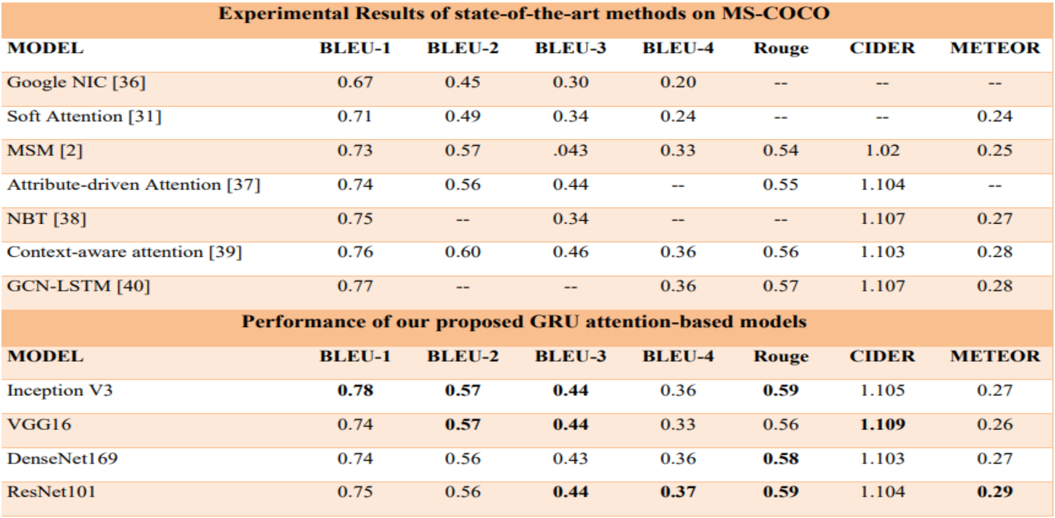
- **Authors: Kelvin Xu ,Ryan Kiros, Kyunghyun Cho, Aaron Courville, ,Ruslan Salakhutdinov, Richard S. Zemel, ,Yoshua Bengio**
- Dataset used Flickr9k, Flickr30k and MS COCO.
- Done **using conditional random field** (CRF) prediction image tag to generate a **natural language description, Attention mechanism** techniques
-



Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)†Σ	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)°	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC†°Σ	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46

4) A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism

- **Authors: Rashid khana , M Shujah Islama , Khadija Kanwala , Mansoor Iqbal, Md. Imran Hossaina & Zhongfu Ye**
- **Dataset used MS COCO.**
- combined the Bahdanau attention model with GRU to allow learning to be focused on a specific portion of the image in order to improve the performance.
- The model's performance was compared to that of four pre-trained CNNs: InceptionV3, DenseNet169, ResNet101, and VGG16.

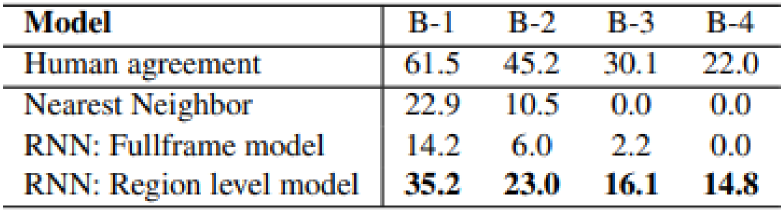


Experimental Results of state-of-the-art methods on MS-COCO							
MODEL	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Rouge**	**CIDER**	**METEOR**
Google NIC [36]	0.67	0.45	0.30	0.20	--	--	--
Soft Attention [31]	0.71	0.49	0.34	0.24	--	--	0.24
MSM [2]	0.73	0.57	.043	0.33	0.54	1.02	0.25
Attribute-driven Attention [37]	0.74	0.56	0.44	--	0.55	1.104	--
NBT [38]	0.75	--	0.34	--	--	1.107	0.27
Context-aware attention [39]	0.76	0.60	0.46	0.36	0.56	1.103	0.28
GCN-LSTM [40]	0.77	--	--	0.36	0.57	1.107	0.28
Performance of our proposed GRU attention-based models							
MODEL	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Rouge**	**CIDER**	**METEOR**
Inception V3	**0.78**	**0.57**	**0.44**	0.36	**0.59**	1.105	0.27
VGG16	0.74	**0.57**	**0.44**	0.33	0.56	**1.109**	0.26
DenseNet169	0.74	0.56	0.43	0.36	**0.58**	1.103	0.27
ResNet101	0.75	0.56	**0.44**	**0.37**	**0.59**	1.104	**0.29**

5) Deep Visual-Semantic Alignments for Generating Image Descriptions

- **Authors: Andrej Karpathy Li Fei-Fei**
- **Dataset used** Flickr8K, Flickr30K and MSCOCO datasets.

- combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions.



Model	B-1	B-2	B-3	B-4
Human agreement	61.5	45.2	30.1	22.0
Nearest Neighbor	22.9	10.5	0.0	0.0
RNN: Fullframe model	14.2	6.0	2.2	0.0
RNN: Region level model	**35.2**	**23.0**	**16.1**	**14.8**