SURVEY ON IMAGE CAPTIONING USING CNN AND LSTM WITH NETWORK COMPRESSION

***Lalit T. Choudhary, **Prof. Vina M. Lomte, *Janhvi Y. Bobhate, *Bhagyashree N. Mulay, *Shweta A. Singh**

**Bachelor of Computer Engineering, **Head of Department of Computer Engineering*

RMD Sinhgad School of Engineering, Warje, Pune - 411041, Maharashtra, India.

ABSTRACT

For the past few years, neural networks are maturing, and the application domain for the neural network is increasing. Due to the rise in unstructured Image-based data, there's a need for understanding data based upon visual features, not on textual data. Image Captioning in Deep learning is the process to understand different objects in the image, try to build a relation between those objects, and give a sentence that is semantically and syntactically correct. To serve this purpose, Image Captioning uses the combined architecture of Computer Vision and NLP-based Networks as Encoder-Decoder Architecture. In this Survey paper, we discuss the paper[1], which proposed to use Efficient CNN-LSTM based Network using Network Compression. The Network uses CNN-based VGG-16 as encoder and LSTM as decoder network. Network techniques are used such as Quantization and Pruning, to reduce model size up to 73.1% and reduce inference time up to 71.3% and increase BLEU score to 7.7% as compared to uncompressed network.

Keywords : *Computer Vision, Deep Learning, Encoder- Decoder Network, Image Captioning, Image Processing, LSTM, Natural Language Processing(NLP), VGG-16.*

INTRODUCTION:

In recent years, sources for image data have been increased exponentially. Images and video content has been the primary source for the transfer of information in this modern world. And most of this data is unstructured or no metadata is attached with that image, so it's become harder for any architecture to understand the context of the image and process based upon the image. These days, we require some automated tools to process this visual data and extract necessary information from this data. Usually, where comes the part to understand an image, it depends upon textual data such as comments, titles, and metadata. So the previous work depends upon certain activities, so to remove this dependency we use the Image Captioning process.

Image Captioning in Deep learning is the process to understand different objects in the image, try to build a relation between those objects, and give a sentence that is semantically and syntactically correct. Image captioning comprises the architecture based upon Convolutional Neural Network (CNN) and Natural Language Processing (NLP) as Encoder-Decoder Network. In this paper, CNN based VGG-16[2] or ResNet50[3] model is used as Encoder, and NLP based Long Short-Term Memory[4] Network is used as Decoder.

But due to the increase in the complexity of the networks, it has created a challenge to deploy

these networks in real-time devices because of computational

cost, computation support, and execution time. In this paper, the Image Captioning model is post-processed with the techniques called Quantization[1] and Pruning[1] which reduce the inference time and model size and increase the BLEU Score.

Figure 1: Generating Caption for Images

PROPOSED ARCHITECTURE:

In this System, VGG-16 is used as an encoder, whereas ResNet-50 can also be used instead of VGG-16, and LSTM is used as a decoder network and techniques like quantization and pruning are used for optimization.

VGG-16 accepts images in 224 x 224 x 3 dimensions, so the image is pre-processed into a particular dimension. After pre-processing the image is passed to the encoder network. Encoder network encodes the data and extracts the necessary information from the image and after encoding the data, it passes through Quantization and Feature extraction to reduce the model size and inference time. This encoding data is now passed to the Decoding Side of the Network as the last layer of post-processing is connected to the decoder network.

On the Decoder side, Long Short-Term Memory Network is used, where the encoding is mapped to the pretrained LSTM network and it forms the relation among the encodings and forms a sentence that is syntactically and semantically correct. This sentence is given as output as Image Caption. Below figure 2 shows the Architecture for the Image Captioning Model using CNN - LSTM Based Encoder Decoder Network with Network Compression.

The model is trained using LSTM decoder and VGG16 or ResNet50 as an encoder on the flickr8k dataset. Flickr 8k Dataset consists of 8000 images, with 5 different captions for each image. They have designed a compression architecture, by studying the effects of different compression architectures on the model that achieves a 73.1% reduction in model size, 71.3% reduction in inference time and a 7.7% increase in BLEU score over its uncompressed counterpart.

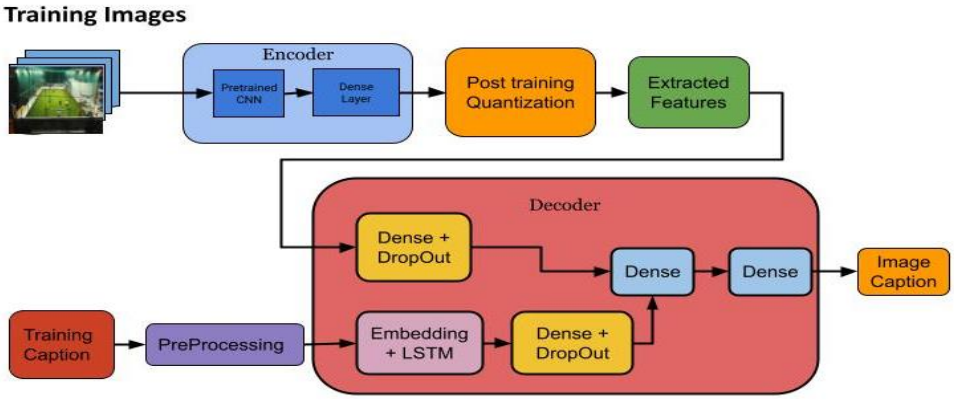

Figure 2 : System Architecture "Image Captioning Model using CNN - LSTM Based Encoder Decoder Network with Network Compression

ARCHITECTURE COMPONENTS:

The Encoder - Decoder Network are individually trained as separate parts, so let's discuss them individually.

A) **VGG-16 Network[2] (Encoder):** In image Captioning, VGG-16 has been used as an encoder which represents CNN model, image is given to this model, it extracts the visual features and gives output as a dictionary of tags it detects and localizes in the image. K.Simonyan and A.Zisserman from the University of Oxford proposed VGG16 convolutional neural network model in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". The model succeeded in achieving a 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images that belong to 1000 classes. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 in the first and 5in the second convolutional layer) with multiple 3×3 kernel-sized filters one after another. Input to cov1 layer = 224 x 224 RGB image.

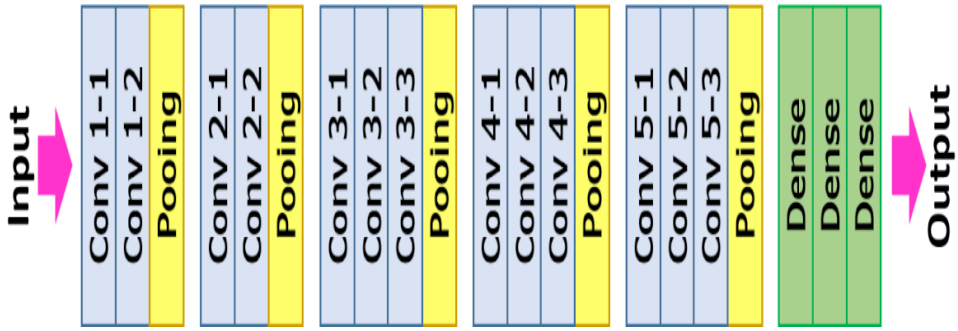

Figure 3: VGG-16 Network Architecture

B) **LSTM Network[4] (Decoder):** The decoder model is LSTM (Long Short Term Memory). The

core concept of LSTM 's are the cell state, and it's various gates. Cell states store the information throughout the network and gates control the logic of the information.The decoder produces captions for an image based on the extracted features of the image from the encoder. We use the the processed captions and extracted features of the training images from the encoder model to train the decoder.Using standard pre-processing procedures like converting characters to lower case, removing punctuation and digits, processing of text is done after which the vocabulary is created followed by its tokenization. We design a multi-input decoder model to process the extracted features and the texts to produce captions. The Feature extractor part has a dense layer with 256 neurons and a dropout layer, the text extractor part pre-process the training captions and is followed by an embedding layer and an LSTM layer of 256 neurons. The decoder layer merges the outputs from the feature and the text extractor layers, followed by two dense layers, one with as many neurons as the vocabulary size and the other with 256 neurons.

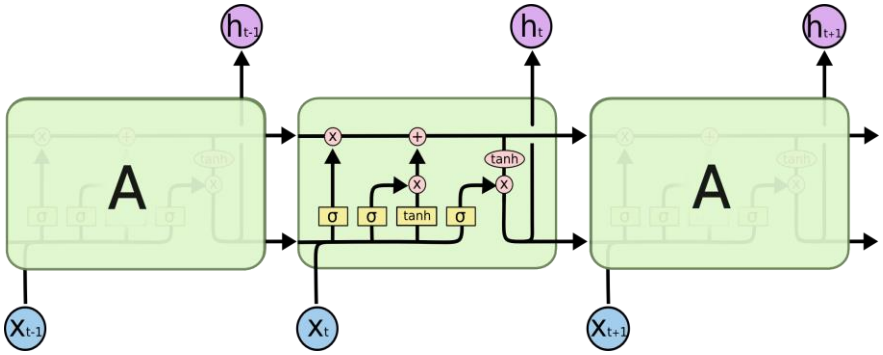

Figure 4: LSTM Network Cell Architecture

C) **Quantization[1] :**For deep learning, quantization is the process of approximating a neural network that uses floating-point numbers by a neural network of low bit width numbers. This dramatically reduces both the computational cost and memory requirement of using neural networks. Quantization is about learning a reduced bit precision without compromising

$$\text{Value}_{float32} = \text{scale} \times (\text{Value}_{int8} - \text{Zeropoint})$$

network performance and it is recognized as one of the most effective approaches to satisfy low memory requirements of resource constrained devices.

D) **Pruning[1] :** Pruning is a technique in deep learning that aids in the development of smaller and more efficient neural networks. It's a model optimization technique. Pruning is a method for inference to efficiently produce models smaller in size, more memory-efficient, more power-efficient and faster at inference with minimal loss in accuracy.

$$s_t = s_f + (s_i - s_f)\left(1 - \frac{t - t_0}{n\Delta t}\right)^3$$

LITERATURE SURVEY :

Sr no	Published Year	Published By	Research Topic	Research Gap	Access Parameter	Outcomes
1	2020	Li Wang, Zechen Bai, Yonghua Zhang, Hongtao Lu	Show, Recall, and Tell: Image Captioning with Recall Mechanism[5]	This mechanism showed lesser results compared to other models.	Recall mechanisms to imitate the way humans conduct captioning. There are three parts: recall unit, semantic guide (SG) and recalled-word slot (RWS).	The experiments prove that the recall mechanism can effectively employ recalled information to improve the quality of generated captions.
2	2020	Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum	Clevrer: Collision events for video representation and reasoning.[6]	There is a decrease of model performance on tasks that require long-term dynamics prediction like the counterfactual questions require a better dynamics model capable of generating more stable and accurate trajectories.	The CLEVRER design works on two guidelines: 1. The listed tasks should focus on logic reasoning in the temporal as well as causal domain.2. The dataset should be well-annotated and fully controlled.	The introduced set of benchmark tasks to better facilitate the research in this area also believe video understanding and reasoning.
3	2019	Hanzhang Wang , Hanli Wang* , Kaisheng Xu	Swell-and-Shrink: Decomposing Image Captioning by Transformation and Summarization[7]	Once generating undesired descriptions, it usually relies on the method of trial and error to diagnose and improve model performance.	The proposed method to redefine image captioning as a compositional task which consists of two separated modules: modality transformation and text compression.	It shows better interpretability on analyzing the process of language formation with regional visual information.
4	2019	Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares	Image Captioning: Transforming Objects into Word [8]	It needs to incorporate geometric attention in our decoder cross-attention layers between objects and words.	Building upon the bottom-up and top down image captioning approach, the proposed Transformer encodes 2D position and size relationships between detected objects in images.	In this paper they have introduced the Object Relation Transformer, that builds upon this approach by explicitly incorporating information about the spatial relationship between input detected objects through geometric attention.
5	2019	Jia Huei Tan, Chee Seng Chan, Joon Huang Chuah	COMIC: Towards A Compact Image Captioning	Future work is to study the impact of different encoders and to train the radix encoding	Approach is made to deal with the problem of compactness of image captioning models.	The outcome of the COMIC experiment shows that overall performance was not affected despite a

			Model with Attention [9]	models in a greedy decoding setting .		reduction of 33×-99× in the vocabulary size.
6	2019	J. H. Tan, C. S. Chan, and J. H. Chuah	Image captioning with sparse recurrent neural network[10]	Further investigation is to be made for the generalisation capability of end-to-end pruning when applied on Transformer models and also like to extend our method to other CV and NLP tasks.	An one-shot end-to-end pruning method is proposed to produce very sparse image captioning decoders (up to 97.5% sparsity) while maintaining good performance relative to the dense baseline model as well as competing methods.	An end-to-end pruning method is proposed that performs considerably better than competing methods at maintaining captioning performance while maximising compression rate.
7	2019	Y. Shangguan, J. Li, L. Qiao, R. Alvarez, and I. McGraw	Optimizing speech recognition for the edge[11]	Attempt to reduce the accuracy losses for the future although they do not compound	Create speech recognizers that are both small and highly accurate, fast enough to execute on typical mobile devices.	A comprehensive set of optimizations is presented spanning from more efficient neural network building blocks to the elimination, and reduction in precision, of neural network parameters and computations.
8	2018	X. Dai, H. Yin, and N. K. Jha	Grow and prune compact, fast, and accurate LSTMs.[12]	To improve the effectiveness of the model future work required.	Hidden-layerLSTM (HLSTM) is added also it increases accuracy. The grow-and-prune (GP) training is used to regulate the hidden layers.	H-LSTM and GP training is performed to learn effectiveness of LSTMs..
9	2018	K. Shi and K. Yu	Structured word embedding for low memory neural network language model[13]	Future work is intended for the effectiveness of the approach and also merging with other compression techniques.	The memory is saved with the help of quantization based structured framework.	Quantization based structured embedding perspective is used which decreases memory utilisation.
10	2018	Marc Tanti, Albert Gatt, Kenneth P. Camilleri	Where to put the Image in an Image Caption	Future scope is to explore the analysis performed on the	To study the inference for caption generator	Estimation of various architectural model for image caption

			Generator[14]	other applications of conditioned neural language models.	architectures using various ways based on visual information.	generation and retrieval has been evaluated systematically.Also studied the 'inject' and 'merge' architectures.
11	2018	CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL	Image captioning with deep Bidirection-al LSTMs and multi-task learning,[15]	Future scope is to scrutinize the multilingual caption generation problem ,video captioning and text recognition.	Combination of Bi-LSTM and CNN helped the model for visual-language communication.	It proved that multi-task learning of Bi-LSTM increased model generality, the effectiveness, and robustness of the proposed models were evaluated on two different tasks: image captioning and image-sentence retrieval.
12	2018	Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama and Kevin Murphy	Improved Image Captioning via Policy Gradient optimization of SPIDEr[16]	Limitation is these metrics do not correlate well with human judgement	Proposed a new policy gradient method and a new metric, SPIDEr which identifies criteria for a good image captioning metric	They proposed efficient policy gradient method resulted in optimizing a variety of captioning metrics and algorithms that produced qualitative results.
13	2017	Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Chua	SCA-CNN: Spatial and Channel-wise Attention in Convolutiona l Networks for Image-Captioning[17]	Future work would be to bring temporal attention , features in different video frames for video captioning and to increase the number of attentive layers without overfitting.	SCA-CNN actively regulates the sentence generation part.	SCA-CNN performs state-of-the-art visual attention-based image captioning methods and helps in better understanding of where and what the attention looks like in a CNN that evolves during sentence generation.
14	2017	Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv and Li-Jia Li.	Deep Reinforcemen t Learning-based Image Captioning with Embedding Reward[18]	To improve the network architectures and investigate the other embedding measures.	An actor-critic reinforcement learning algorithm is introduced which is driven by visual-semantic embedding	A novel decision-making framework is used for image captioning which achieves state-of-the-art performance on standard benchmark.

15	2017	X. Dai, H. Yin, and N. Jh	NeST: A neural network synthesis tool based on a grow-and prune paradigm[19]	Attempts would be made in future to improve the architecture so that it arrives at a compact NN with high accuracy by adjusting the architecture.	NeST initially starts with a seed NN architecture and iteratively tunes the architecture with gradient-based growth and magnitude based pruning of neurons and connections.	Thus presented NeST tool which will give precise NNs.
16	2017	A. Krizhevsky, I. Sutskever, and G. E. Hinton.	ImageNet classification with deep convolutional neural networks[20]	Use very large and deep convolutional nets on video sequences for future purpose where the temporal structure provides very helpful information, that is, missing or far less obvious in static images.	Collect larger datasets to improve the performance of machine learning methods, learn more powerful models, and use better techniques for preventing overfitting.	Thus,deep CNN is efficient of accomplishing results on a challenging dataset using purely supervised learning layer is removed.
17	2016	Kelvin Xu,Jimmy Lei Ba,Ryan Kiros, Kyunghyun Cho, Aaron Courville,Ruslan Salakhutdinov,Richard S. Zemel,Yoshua Bengio	Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[21]	For future, work would be done in using visual attention and also expects the modularity of the encoder-decoder approach.	Used the state-of-the art performance on the datasetsand also trained the model in deterministic manner.	Using the BLEU and METEOR metric, proposed an attention based approach that gives state of the art performance.
18	2014	Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik.	Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections[22]	In the future, they would gain more insight into what makes SAE effective	A new algorithm is introduce that can successfully transfer knowledge from millions of weakly annotated images to improve the accuracy of retrieval-based image description	Thus, established new approach named Stacked Auxiliary Embedding which helps to enhance the precision of retrieval-based image description.

Table 1. Literature Survey

ALGORITHMIC/ACCURACY SURVEY:

Sr no	Paper Title	Publication Details	Algorithm / Model Architecture	Accuracy	Research Gap
1	Show and Tell: A Neural Image Caption Generator	IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL.XX, NO. XX, MONTH 2016[23]	CNN - RNN Based Encoder - Decoder Network	On COCO Dataset, model achieve a BLEU-4 of 27.7	This research limits to particular datasets, so it needs to add a Dataset and Achieve State of Art Performance.
2	Bottom-up and top-down attention for image captioning and visual question answering	P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , page 6077-6086. IEEE Computer Society, (2018)[24]	Bottom-up and top-down attention model	On the MSCOCO test server establish a new state-of-the-art for the task, achieving CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5 and 36.9, respectively	This work more closely at a higher level unifies tasks involving visual and linguistic understanding with recent progress in object detection.
3	Image Captioning with semantic attention	You, Quanzeng & Jin, Hailin & Wang, Zhaowen & Fang, Chen & Luo, Jiebo. (2016). Image Captioning with Semantic Attention. 4651-4659. 10.1109/CVPR.2016. 503.[25]	Input Attention Model and Output Attention Model.	On the MSCOCO model, BLEU-4 scores of 0.709, 0.537, 0.402, 0 .304 resp. and on the Flickr30k model BLEU-4 scores of 0.647, 0.460, 0.324, 0.230 respectively.	This paper explores new models for our proposed semantic attention mechanism.as well as to experiment with phrase-based visual attributes with its distributed representations.

Table 2 : Algorithmic/ Accuracy Survey

CONCLUSION :

The main purpose of this study was to study real time based solutions for Image Captioning. In this study we presented an Efficient CNN- LSTM based compressive Neural Network which uses VGG-16 and LSTM network as Encoder - Decoder Network and discussed optimization techniques like Quantization And Pruning technique to reduce inference runtime and model size and it is observed that after quantization the performance for the model is increased on Flickr 8k Dataset. The Network is better compared only because of optimization techniques, so in spite of optimization techniques it has achieved good results on Dataset as compared to unoptimized network. Limitation is the scope of the network, so in future improvement will be increasing the scope of tha dataset and achieving same or higher performance from the network.

REFERENCES :

[1] Harshit Rampal and Aman Mohanty, "*Efficient CNN-LSTM based Image Captioning using Neural Network Compression", rXiv:2012.09708 [cs.CV]*

[2] S. Liu and W. Deng, *"Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 2015, pp. 730-734, doi: 10.1109/ACPR.2015.7486599.*

[3] K. He, X. Zhang, S. Ren and J. Sun, *"Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.*

[4] Yu Wang, *"A new concept using LSTM Neural Networks for dynamic system identification," 2017 American Control Conference (ACC), Seattle, WA, 2017, pp. 5324-5329, doi: 10.23919/ACC.2017.7963782.*

[5] Wang, Li & Bai, Zechen & Zhang, Yonghua & Lu, Hongtao. (2020). *Show, Recall, and Tell: Image Captioning with Recall Mechanism. Proceedings of the AAAI Conference on Artificial Intelligence. 34. 12176-12183. 10.1609/aaai.v34i07.6898.*

[6] Yi, Kexin, Chuang Gan, Yunzhu Li, P. Kohli, Jiajun Wu, A. Torralba and J. Tenenbaum. *"CLEVRER: CoLlision Events for Video REpresentation and Reasoning." ArXiv abs/1910.01442 (2020): n. pag.*

[7] Wang, Hanzhang and Wang, Hanli and Xu, Kaisheng, *Swell-and-Shrink: Decomposing Image Captioning by Transformation and Summarization, International Joint Conferences on Artificial Intelligence Organization, 2019, 5226--5232, doi: 10.24963/ijcai.2019/726.*

[8] Herdade, Simao et al. *"Image Captioning: Transforming Objects into Words." NeurIPS (2019).*

[9] Tan, Jia Huei & Chan, Chee Seng & Chuah, Joon Huang. (2019). *COMIC: Towards A Compact Image Captioning Model with Attention. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2019.2904878.*

[10] Tan, Jia Huei & Chan, Chee Seng & Chuah, Joon Huang. (2019). *Image Captioning with Sparse Recurrent Neural Network.*

[11] Shangguan, Yuan & Li, Jian & Qiao, Liang & Alvarez, Raziel & McGraw, Ian. (2019). *Optimizing Speech Recognition For The Edge.*

[12] Dai, Xiaoliang & Yin, Hongxu & Jha, N.K.. (2019). *Grow and Prune Compact, Fast, and Accurate LSTMs. IEEE Transactions on Computers. PP. 1-1. 10.1109/TC.2019.2954495.*

[13] Shi, Kaiyu and K. Yu. *"Structured Word Embedding for Low Memory Neural Network Language Model." INTERSPEECH (2018).*

[14] Tanti, Marc & Gatt, Albert & Camilleri, Kenneth. (2017). *Where to put the Image in an Image Caption Generator. Natural Language Engineering. 24. 10.1017/S1351324918000098.*

[15] Wang, Cheng & Yang, Haojin & Meinel, Christoph. (2018). *Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. ACM Transactions on Multimedia Computing, Communications, and Applications. 14. 1-20. 10.1145/3115432.*

[16] Liu, Siqi & Zhu, Zhenhai & Ye, Ning & Guadarrama, Sergio & Murphy, Kevin. (2017). *Improved Image Captioning via Policy Gradient optimization of SPIDEr. 873-881. 10.1109/ICCV.2017.100.*

[17] Chen, Long & Zhang, Hanwang & Xiao, Jun & Nie, Liqiang & Shao, Jian & Liu, Wei & Chua, Tat-Seng. (2017). *SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. 6298-6306. 10.1109/CVPR.2017.667.*

[18] Z. Ren, X. Wang, N. Zhang, X. Lv and L. Li, *"Deep Reinforcement Learning-Based Image Captioning with Embedding Reward," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1151-1159, doi: 10.1109/CVPR.2017.128.*

[19] Dai, Xiaoliang & Yin, Hongxu & Jha, N.K.. (2017). *NeST: A Neural Network Synthesis Tool Based on a Grow-and-Prune Paradigm. IEEE Transactions on Computers. PP. 10.1109/TC.2019.2914438.*

[20] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). *ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.*

[21] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. *Show, attend and tell: neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 2048–2057.*

[22] Gong, Yunchao & Wang, Liwei & Hodosh, Micah & Hockenmaier, Julia & Lazebnik, Svetlana. (2014). *Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. 529-545. 10.1007/978-3-319-10593-2_35.*

[23] Vinyals, Oriol & Toshev, Alexander & Bengio, Samy & Erhan, Dumitru. (2015). *Show and tell: A neural image caption generator. 3156-3164. 10.1109/CVPR.2015.7298935.*

[24] Anderson, Peter & He, Xiaodong & Buehler, Chris & Teney, Damien & Johnson, Mark & Gould, Stephen & Zhang, Lei. (2018). *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 6077-6086. 10.1109/CVPR.2018.00636.*

[25] You, Quanzeng & Jin, Hailin & Wang, Zhaowen & Fang, Chen & Luo, Jiebo. (2016). *Image Captioning with Semantic Attention. 4651-4659. 10.1109/CVPR.2016.503.*