**LITERATURE SURVEY**

# 1) An Empirical Study of Language CNN for Image Captioning

**Authors:** [Jiuxiang Gu](https://arxiv.org/search/cs?searchtype=author&query=Gu%2C+J)**,** [Gang Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang%2C+G), [Jianfei Cai](https://arxiv.org/search/cs?searchtype=author&query=Cai%2C+J) and [Tsuhan Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen%2C+T)

Language Models based on recurrent neural networks have dominated recent image caption generation tasks. In this paper, we introduce a Language CNN model which is suitable for statistical language modeling tasks and shows competitive performance in image captioning. In contrast to previous models which predict next word based on one previous word and hidden state, our language CNN is fed with all the previous words and can model the long-range dependencies of history words, which are critical for image captioning. The effectiveness of our approach is validated on two datasets MS COCO and Flickr30K. Our extensive experimental results show that our method outperforms the vanilla recurrent neural network based language models and is competitive with the state-of-the-art methods.

# 2.) Convolutional Image Captioning

**AUTHORS:** [Jyoti Aneja](https://arxiv.org/search/cs?searchtype=author&query=Aneja%2C+J),[Aditya Deshpande](https://arxiv.org/search/cs?searchtype=author&query=Deshpande%2C+A), and  [Alexander Schwing](https://arxiv.org/search/cs?searchtype=author&query=Schwing%2C+A)

Gippsland School of Information Technology, Monash University, Churchil

Image captioning is an important but challenging task, applicable to virtual assistants, editing tools, image indexing, and support of the disabled. Its challenges are due to the variability and ambiguity of possible image descriptions. In recent years significant progress has been made in image captioning, using Recurrent Neural Networks powered by long-short-term-memory (LSTM) units. Despite mitigating the vanishing gradient problem, and despite their compelling ability to memorize dependencies, LSTM units are complex and inherently sequential across time. To address this issue, recent work has shown benefits of convolutional networks for machine translation and conditional image generation. Inspired by their success, in this paper, we develop a convolutional image captioning technique. We demonstrate its efficacy on the challenging MSCOCO dataset and demonstrate performance on par with the baseline, while having a faster training time per number of parameters. We also perform a detailed analysis, providing compelling reasons in favor of convolutional language generation approaches.

# 3. ) Image captioning with deep LSTM based on sequential residual

# 

**AUTHORS :** [Hanli Wang](https://ieeexplore.ieee.org/author/37292564900) [,](https://arxiv.org/search/cs?searchtype=author&query=Roth%2C+P)[Pengjie Tang](https://ieeexplore.ieee.org/author/37086172987) and [Kaisheng Xu](https://ieeexplore.ieee.org/author/37086529906)

Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. In consideration of the problem that language models are always shallow in modern image caption frameworks, a deep residual recurrent neural network is proposed in this work with the following two contributions. First, an easy-to-train deep stacked Long Short Term Memory (LSTM) language model is designed to learn the residual function of output distributions by adding identity mappings to multi-layer LSTMs. Second, in order to overcome the over-fitting problem caused by larger-scale parameters in deeper LSTM networks, a novel temporal Dropout method is proposed into LSTM. The experimental results on the benchmark MSCOCO and Flickr30K datasets demonstrate that the proposed model achieves the state-of-the-art performances with 101.1 in CIDEr on MSCOCO and 22.9 in B-4 on Flickr30K, respectively.

# 4. ) Automatic Image Captioning Using Convolution Neural Networks and LSTM

**AUTHORS:** [**rajendran subash**](https://www.researchgate.net/profile/Rajendran-Subash?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIiwicGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIn19)**,** [**Jebakumar Rethnaraj**](https://www.researchgate.net/profile/Jebakumar-Rethnaraj?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIiwicGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIn19)**,** [**Yash Kamdar**](https://www.researchgate.net/scientific-contributions/Yash-Kamdar-2166691325?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIiwicGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIn19)

PC vision has turned out to be universal in our general public, with applications in a few fields. Given a lot of pictures, with its inscription, make a prescient model which produces regular, inventive, and intriguing subtitles for the concealed picture. A speedy look at a picture is adequate for a human to call attention to and portray a monstrous measure of insights regarding the visual scene. To rearrange the current issue of producing inscriptions for pictures by making a model which would give exact subtitles to these pictures which can be additionally utilized in other helpful applications and use cases. Be that as it may, this momentous capacity has ended up being a tricky errand for our visual acknowledgment models. Most of the past research in scene acknowledgment has concentrated on naming pictures with a predetermined arrangement of visual classifications and extraordinary advancement has been accom- plished in these undertakings. For a question picture, the past strategies recover pertinent hopeful normal language states by outwardly contrasting the inquiry picture with database pictures. In any case, while shut vocabularies of visual ideas comprise a helpful demonstrating suspicion, they are boundlessly prohibitive when contrasted with the colossal measure of rich depictions that a human can form. These methodologies forced a breaking point on the assortment of inscriptions produced. The model ought to be exempt of suppositions regarding explicit pre decided formats, standards or classes and rather depend on figuring out how to create sentences from the preparation information. The model proposed utilizes Convolution Neural Networks which help to separate highlights of the picture whose subtitle is to be created and afterward by utilizing a probabilistic methodology and Natural Language Processing Techniques reasonable sentences are framed and inscriptions are produced.