IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101273

Review Article on Image Captioning

Harsh Mehta[1] , Vipul Jain[2] , Shivani Patel[3] , Kriti Banthia[4] , Jitender Jaiswal[5]

Faculty of Engineering and Technology, Jain University, Bengaluru, India[1-5]

Abstract: Image captioning is a task that tries to generate captions for the given photographs by combing computer vision and natural language processing. It's a two-step process in which precise image recognition and appropriate syntactic and semantic language comprehension. Due to the rising amount of information accessible on this subject, keeping up with the newest research and findings in the fieldof picture captioning is becoming increasingly difficult.
Current research in the field is mostly focused on deep learning-based methods, with attention mechanisms, deep reinforcement, and adversarial learning appearing to be at the forefront. In this paper we will go through various research papers which focuses on deep learning models and uses COCO dataset or Flicker dataset.

INTRODUCTION

People communicate through language and they often use this language to describe the world in frontof them. Images are also used to describe the world in pictorial representation.
But for computers it is all the same. Image captioning is field where we can use computer to describe images.Captioning of images have many uses from automatic picture indexing to assistive technologies.
But generation of meaningful descriptions from images is a very challenging task. This task of automatically generating captions and describing the image is significantly harder than image classification and object recognition.
The description should not only involve the objects in the image but also relation between the objects with their attributes and activities shown The employment of attention mechanisms is one of the approaches that plays a critical part in picture captioning today.[1]

RELATED WORK

As mentioned in the review paper[2], the authors presented a comprehensive review of the state-of- the-art deep learning-based image captioning techniques by late 2018.
The paper gave summary of existing technologies, compared the cons and pros and they also discussed both strength and weakness of datasets and evaluation metrics.
The paper published in 2019[3] compared different image captioning models from 2016 to 2019 on Flickr30k and MSCOCO dataset. An investigation was done on different feature extractors including
AlexNet, VGG-16 Net, ResNet, GoogleNet with all the nine Inception models, and DenseNet. In addition, language models were covered such as LSTM, RNN, CNN, GRU and TPGN. This comparisonwas evaluated on various metrics like BLEU(1-4), CIDEr and METEOR.
Paper Published in 2020[4] stated that CNN-LSTM outperformed CNN-RNN models, they also performed a comparison on models from 2016 to 2019 and evaluated on BLEU(1-4) metrics.
Image captioning remains an active research area, and new methodologies keep being publishedup until this moment.

METHODS

hLSTMat

Most decoders apply attention mechanism to both visual words('book','reading') and non visual words('a', 'the') but these non visual words can mislead the model and also the hierarchy of LSTMs enables more complex representation of visual data, capturing information at different scales.To address these issues, they proposed a hierarchical LSTM with adaptive attention (hLSTMat) approach[10] forimage and video captioning.
The proposed framework utilizes the spatial or temporal attention for selecting specific regions or frames to predict the related words, while the adaptive attention is for deciding whether to depend onthe visual information or the language context information. Also, a hierarchical LSTMs is designed to simultaneously consider both low-level visual information and high-level language context information to support the caption generation.
The framework of our proposed hLSTMat for visual captioning. Given an input image or video, an encoder is firstly applied to extract the features. Then hierarchical LSTM with adaptive attention component plays the role of an decoder, by using the hierarchical LSTM to extract different level ofinformation, and an adaptive attention to decide whether to depend on the visual information or the language context information. The losses are defined on the generated captions and the ground truth to guide the learning of network parameters.

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 12, December 2021
DOI: 10.17148/IJARCCE.2021.101273

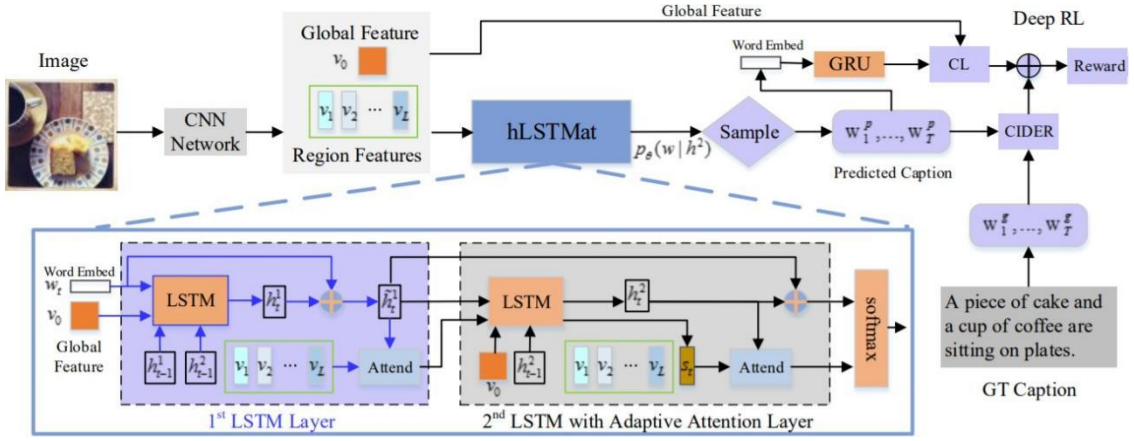

Fig. 1: model[10] uses residual shortcut connection to improve information flow through two LSTMs. And adaptive attention is applied to calculate weights of features when predicting new word

UpDown

Most method which uses visual attention are of top-down kind, the issue with this approach is that there is no deliberation as which region will receive attention.

Introduction of Updown[5], a model that joins an entirely visual bottom-up mechanism and a task-specific context top-down one. The first part proposes which part deems salient and the latter used to context to compute an attention distribution over them.

The bottom-up mechanism employs the Faster R-CNN object detection model, responsible for recognizing class objects. The top-down mechanism uses a visual attention LSTM and a language one. The attention LSTM is fed the previous language LSTM outputs, The word generated at time t-1 and mean-pooled image features to decide which regions should receive attention.

OSCAR

Vision-language pre-training (VLP) is widely used for learning cross-modal representations. It suffers, however, from 2 issues: a difficulty in differentiating features due to overlap of image regions and alack of alignment caption words.

The overcome these issues OSCAR[6] uses object tags as anchor points. The use of three inputs composed of image region feature, object tags and word sequence. By this if one of the three input isnoisy or unclear then the other inputs can be complete the information It is therefore simple to makethe alignments,because the most important elements in the image appear in the matching caption.

OSCAR detects object tags using Faster R-CNN and presents a 2-view perspective.

(1.) Dictionary view with a linguistic semantic space compressing the tags and captions tokens

(2.) A modality-view that consist of an image modality containing image features, tags and caption tokens

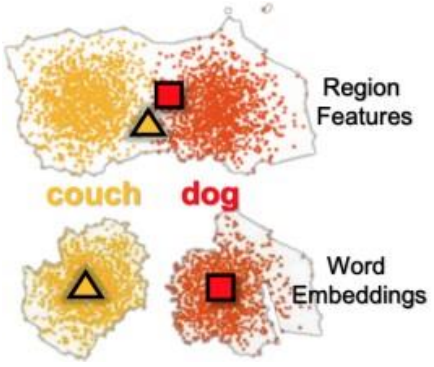

Fig. 2: The semantic space used by OSCAR[6]. In the example of a dog sitting on a couch, ”couch” and ”dog” are close in region features since they’re roughly in the same area of the image, but they farther apart in word embedding because of their different meanings.

VIVO

VIVO[7] shortened for VIsual VOcabulary pre-training. It creates a visual vocabulary which is a jointembedding space of tags and images region features where vectors are semantically close to objects. After pre-training the vocabulary, the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101273

model is fine-tuned with image-caption pairs using the MS COCO dataset.
VIVO uses a multi-layer Transformer responsible for aligning tags with their corresponding imageregion features. During pre-training, image region features are extracted from the input image usingUpdown's object detector[5] and fed to the Transformer along with a set of pairs of images and tags. In fine-tuning, the model is fed a triplet of image regions, tags and a caption.At inference time,
image region features are extracted from the input image and tags are detected. A caption is then generated one token at a time.

Meta Learning

One of the drawbacks of reinforcement learning is overfitting on the reward function which occurs when the agent finds a way to maximize the score without generating captions of a better quality. When a short caption is generated, common phrases are added to it to make it longer, ending up withunnatural sentence endings such as "a little girl holding a cat in a of a."
[8]introduce meta learning, which is learning a meta model that is able to optimize and adapt to several different tasks. In this case, the model simultaneously optimizes the reward function (reinforcement task) and uses supervision from the ground truth (supervision task) by taking gradient steps in both directions. This guarantees the distinctiveness of the captions and their propositional correctnessand results in sound human-like sentences.

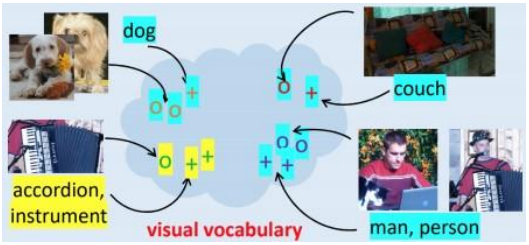

Fig. 3: Visual Vocabulary used by VIVO.[7] Objects that are similar semantically are closer together.o represents regions and + represents tags. Yellow objects and tags are novel .

Conditional GAN-Based Model

To overcome reward hacking, [9] use discriminator networks to decide whether a generated caption is from a human or a machine.
They experimented with two different architectures for discriminator, one of them using a CNN with a fully connected layer and a sigmoid transformation, and the other an RNN (LSTM) with a fullyconnected layer and a softmax. They also experiment with an ensemble of 4 CNNs and 4 RNNs. Forgenerator the used Updown architecture.
The generator and discriminator need to be pre-trained before being alternatively fine-tuned

Evaluation Metrics

To compare the generated captions with the ground truth there are number of evaluation metrics used. Commonly used are CIDEr, SPICE, BLEU and METEOR. Among these most commonly used are CIDEr and SPICE.
CIDEr[11] is an image classification metric that uses term frequency-inverse document frequency to achieve human consensus. SPICE[12] is a new semantic concept-based caption assessment metric based on scene-graph, a graph-based semantic representation.

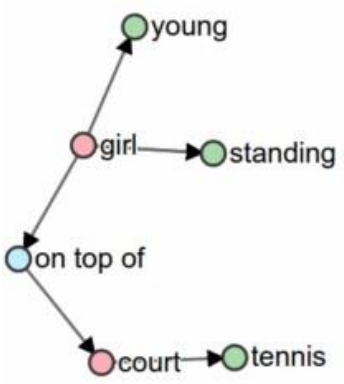

Fig. 4: SPICE scene graph for the caption "A young girl standing on top of a tennis court". The objects are marked red, the relations blue and the attributes green.[11]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101273

Benchmarks

Nocaps benchmark[14] uses images from MSCOCO dataset and open images object detection to introduce novel objects not seen in MSCOCO. It is made up of 166,100 captions that describes 15,100 images. OSCAR, VIVO, UpDown methods are evaluated on nocaps benchmark Karpathy splits[13] are used in the evaluation of the meta learning model, hLSTMat, IC-GAN and UpDown.

RESULTS

MS COCO Karpathy Splits Benchmark

As we can see from the Table:1 UpDown shows an important gain in performance when compared with Resnet baseline. Therefore adding bottom-up attention has an important positive impact on image captioning.

We observe that meta learning receives a CIDEr score of 121.0 and a SPICE score of 21.7. It is the most performant on both evaluation metrics compared to maximizing using the maximum likelihood estimate, reinforcement learning and the MLE+RL maximization.

IC-GAN with Updown/ensemble model has outperformed all other models although hLSTMat model with deliberation process (De) and reinforcement learning (RF) has same score on SPICE metric and its nearly same performance on CIDEr metrics.

Method	CIDEr	SPICE
Resnet Baseline	111.1	20.2
UpDown	120.1	21.4
MLE Maximization	110.2	20.3
RL Maximization	120.4	21.3
MLE+ RL Maximization	119.3	21.2
Meta Learning	121.4	21.7
hLSTMat (DA+De-RF)	111.9	20.5
hLSTMat (DA+De+RF)	125.6	22.3
IC-GAN (Updown/CNN-GAN)	123.2	22.1
IC-GAN (Updown/RNN-GAN)	122.2	22.0
IC-GAN (Updown/ensemble)	125.9	22.3

Table 1: Results of the overall performance on MS COCO Karpathy test split

nocaps Benchmark

The OSCAR model is characterized highly efficient as its uses anchor points making the semantic alignments learning easier. When its on its own it outperforms Updown model and the performance increases tremendously when added with constrained beam search (CBS) and self critical sequentialtraining (SCST).

However OSCAR does not perform as well as VIVO as shown in Table:2. The VIVO+SCST+CBS version shows the highest performance with CIDEr scores that even surpass the human ones.

Method	CIDEr	SPICE
UpDown	55.3	10.1
UpDown + CBS	73.1	11.1
OSCAR	63.8	11.2
OSCAR + CBS	79.3	11.9
OSCAR + CBS + SCST	81.1	11.7
VIVO	81.5	12.2
VIVO + CBS	85.3	12.2
VIVO + CBS + SCST	88.3	12.4
Human	87.1	14.2

Table 2: Evaluation on nocaps validation set

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101273

CONCLUSION

Image Captioning is an active research subject with new methodologies coming up frequently with aims to overcome shortcomings of previous models and with better performance. The ongoing research is focused on deep learning models where attention mechanism is used along with deep reinforcement learning and adversarial learning. The state-of-the-art techniques includes Updown, OSCAR, VIVO, GAN based model. GAN based model is the most performant among all, Updown has the most impact and used along side with nearly all models. We hope this review will provide you a better understanding of existing models.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polo- sukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.

[2] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning, 2018.

[3] R. Stani¯ut˙e and D. ˇSesok. A systematic literature review on ˇimage captioning. Applied Sciences, 9(10):2024, 2019.

[4] M. Chohan, A. Khan, M. Saleem, S. Hassan, A. Ghafoor, and M. Khan. Image captioning using deep learning: A systematic literature review. International Journal of Advanced Computer Science and Applications, 11(5), 2020.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up andtop-down attention for image captioning and VQA. CoRR, abs/1707.07998,2017

[6] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference onComputer Vision, pages 121–137. Springer, 2020

[7] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu. Vivo: Visual vocabulary pre-trainingfor novel object captioning, 2021.

[8] N. Li, Z. Chen, and S. Liu. Meta learning for image captioning. Proceedings of the AAAI Conference on Artificial Intelligence,2019

[9] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju. Improving image captioning with conditional generative adversarial nets. Proceedings of the AAAI Conference on Artificial Intelligence, 33:8142–8150, 2019

[10] Gao, L.; Li, X.; Song, J.; Shen, H.T. Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Trans.Pattern Anal. Mach. Intell. 2019

[11] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

[12] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image cap- tion evaluation. Computer Vision – ECCV 2016 Lecture Notes in Computer Science, page 382–398, 2016

[13] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.

[14] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. CoRR, abs/1812.08658,2018

[15] Ahmed Elhagry, Karima Kadaoui: A Thorough Review on Recent Deep Learning Methodologies for Image Captioning, 2021

[16] R. Stani¯ut˙e and D. ˇSesok. A systematic literature review on image captioning. Applied Sciences, 2019