Image Captioning with a CNN-LSTM Implementation and Comparison to Other Methods

Johannah Cushing

Lewis University, Data Science Program

Saco, ME, USA

[Jkcush03200530@yahoo.com](mailto:Jkcush03200530@yahoo.com); [johannahkcushing@lewisu.edu](mailto:johannahkcushing@lewisu.edu)

Index terms: MsCOCO dataset, image captioning, Flickr dataset, neural networks, convolutional neural networks, long-short-term memory neural networks

Abstract

1. Introduction

Automatic image captioning is the task of producing a natural language description (sentence) to accurately reflect the visual content of an image. This field has seen wide growth in recent years with advances in deep learning architectures, computer vison, and natural language processing. Practical applications include image indexing or retrieval. There are applications for context based image retrieval in many areas such as biomedicine, commerce, education, and web searching [1]. Aiding the visually-impaired by transforming visual signals into information which can be communicated via text-to speech technology is another important application of image captioning.

The development of new annotated datasets as well as new models themselves add to the development of the field. Datasets vary in terms of number of images, size of images, number of captions per image, and how broad or narrow a scope the dataset covers.

Modern image captioning generally uses a deep learning framework are produce novel captions. Current research is improving upon methods have focused both on improving language models, and modifications to various types of architectures, such as the popular CNN-RNN (convolutional neural network – recurrent neural network) or CNN-LSTM (convolutional neural network – long short-term memory) architecture. However, newer architectures are also being used such as GAN (generative adversarial nets). Dense captioning, stylized captions, and especially attention-based mechanisms have improved captioning results on several metrics. Attention based methods focus on a certain part of the *important* concepts, i.e. objects or actions which are integral to constructing an accurate image caption.

The goal of this research is to create an encoder decoder model, test on commonly available datasets for image captioning and compare results to current state of the art methods.

1. Literature Survey
   1. Datasets

Flickr8k and Flickr30k datasets [1] -The Flickr datasets are a collection of iamges from Flickr. The Fli9ckr 8k has 8000 images, predivided into train/test/dev sets, with five captions for each image annotated by humans. The 30k has 30,000 images and the set does not provide any fixed split. There are detectors for common objects, a color classifier, and a bias towards selecting larger objects.

MS-COCO dataset [2]- The MS-COCO dataset was designed to advance object recognition by placing the question of object recognition in the context of the broader question of scene understanding. Images of complex everyday scenes containing common objects in their natural context are used. Objects are labeled using per-instance segmentations to aid in precise object localization. The dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. There is a total of 2.5 million labeled instances in 328k images.

ImageNet dataset[3]. This dataset was built upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500- 1000 clean and full resolution images. There are 3.2 images in total.

[Visual Genome](http://visualgenome.org/) dataset [4]: Visual Genome is a dataset and knowledge base created in an effort to connect structured image concepts to language. The database features detailed visual knowledge base with captioning of 108,077 images. contains over 108K images where each image has an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects. The objects, attributes, relationships, and noun phrases in region descriptions and questions answer pairs are canonicalized to WordNet synsets. Designed to better understand the interactions and relationships between objects in an image.

[Conceptual Captions](https://github.com/google-research-datasets/conceptual-captions), a new dataset consisting of ~3.3 million image/caption pairs that are created by automatically extracting and filtering image caption annotations from billions of web pages [5]. It contains an order of magnitude more images than the widely used MS COCO, with a wider variety of images and image captioning styles.

The Visual Genome Dataset, Instagram Dataset, IAPR TC-12 Dataset, Stock3M, MIT-Adobe FiveK, and FlickrSTyle 10k are also used for image captioning.

* 1. Image Captioning Models

There are many different methods which are used for image captioning. Older methods include Local Binary Patterns, Scale-Invariant Feature Transform, and Histogram of Oriented Gradients where features are extracted from input data and then passed to a classifier such as a support vector machine [1]. However, research of the past several years has focused largely on deep machine learning techniques.

Some deep learning methods of image captioning which are template-based. These can generate grammatically correct captions, templates are predefined and captions are fixed in lengths [1]. Retrieval based methods find the visually similar images with their captions from the training data set. The captions for the query image are selected from the caption pool. Although retrieval based methods produce syntactically correct captions, they are general, not image specific, and cannot create semantically correct captions [1]. Novel captions methods use deep learning techniques and are semantically more accurate than previous approaches.

Most captioning methods use visual space for generating captions. However, some use a multimodal space. A typical multimodal space based method contains a language encode part, a vision part, a multimodal space part, and a language decoder part [1]. In these methods, deep neural networks and a multimodal neural language model are used to learn both image and text jointly in multimodal space. Then, the generation part generates captions. The vision part typically uses a CNN as a feature extractor. The language encoder extracts word features and learns dense feature embedding for each word, forwarding the semantic temporal context to the recurrent layers [1]. The multimodal space maps image features and word features into a common space. The map is then fed to the language decoder and captions are generated.

A popular multimodal model was developed by Mao et al [6]. They use multimodal Recurrent Neural Networks (m-RNN) model. The model contains a language model part, a vision part and a

multimodal part. The language model learns a dense feature embedding for each word in the dictionary and stores the semantic temporal context in recurrent layers. The vision part contains a deep CNN which generates the image representation. The multimodal part connects the language model and the deep CNN by a one-layer representation. The model is learned using a log-likelihood cost function. The temporal context being stored in a recurrent architecture allows for arbitrary context length. Improvements over past models are: a two-layer word embedding system in the m-RNN network structure which learns the word representation more efficiently than the single-layer word embedding. Also, the recurrent layer is not used to store the visual information. The image representation is inputted to the m-RNN model along with every word in the sentence description. The capacity of the RNN is used more efficiently and thus a relatively small dimensional layer can be used. It utilizes of the capacity of the recurrent layer more efficiently.

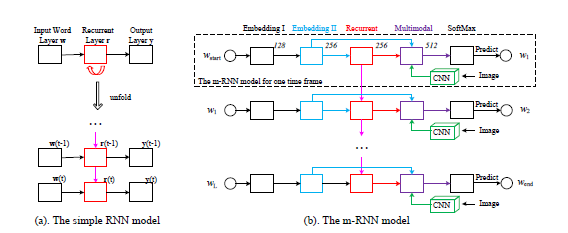


Fig. 1. RNN and mRRN models [6] a) A simple RNN model as compared to b) a m-RNN model used by Mao et al. The inputsof the model are an image and its corresponding sentence descriptions. w1, w2, ..., wL represents the words in a sentence. a start sign wstart and an end sign wend are added to all the training sentences. Themodel estimates the probability distribution of the next word given previous words and the image.It consists of five layers (i.e. two word embedding layers, a recurrent layer, a multimodal layer anda softmax layer) and a deep CNN in each time frame. The number above each layer indicates the dimension of the layer. The weights are shared among all the time frames.

While most models use the whole seen, dense captioning methods generate captions for each region of the scene. DenseCap, proposed by Johnson et al [7] uses a CNN, a dense localization layer, and an LSTM language model, it uses a differential, spatial soft attention mechanism and bilinear interpolation to help back propagate through the network and smoothly select active regions. A problem with dense captioning is that there may be problems recognizing overlapping regions of interest for a particular object.

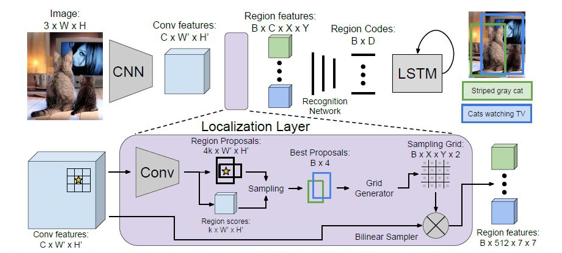


Fig 2. DenseCap model [7].

Captioning models can also be divided into supervised vs. other deep learning models. Supervised based models have been used in tasks such as image classification, object detection and attribute learning. There are also many supervised learning-based image captioning methods- Many of the ones which are commonly heard of—encoder- decoder, compositional architecture, attention-based, semantic concept based, stylized captions, novel object based, and dense image captioning[1] Other deep learning models include .

A common framework is an encoder-decoder architecture. A typical method uses a convolutional neural network (CNN) to obtain the scene type and detect the objects. The output from the CNN is then used by a language model, such as an LSTM to produce captions. Image information is included to the initial state of an LSTM. The next words are generated based on the current time step and the previous hidden state [1], continuing until the end token of a sentence. Recurrent Neural Networks (RNNs) are also commonly used as decoders to convert this representation word-by-word into natural language description of the image. These methods are unable to analyze the image over time while generating the image, and do not consider the spatial aspects of the image that is relevant to the parts of the image captions. [1].

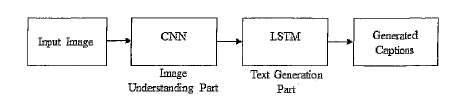


Figure 3. Generic encoder-decoder architecture [1].

LSTM may also face challenges generating longer sentences, since the role of words generated at the beginning becomes weaker and weaker. Jia et al [8] proposed a model called gLSTM (guided LSTM) which is an extension of the long short term memory (LSTM) model In particular, it adds semantic information extracted from the image as extra input to each unit of the LSTM block, with the aim of guiding the model towards solutions that are more tightly coupled to the image content. Unidirectional LSTM cannot generate contextually well-formed captions (Hossain). Wang et al. proposed a deep bidirectional LSTM based method for image captioning which can produce contextually and semantically richer captions. The architecture consist of a CNN and two separate LSTM networks which can utilize both past and future context information to learn long term visual language interactions.

There are compositional architecture-based methods [1]. These are composed of several independent building blocks. Here, a CNN is used to obtain image features. Then visual concepts or attributes are obtained from visual features. Multiple captions are generated by a language model. To generate a final caption, these candidate captions are re-ranked using a multimodal similarity model. An example is proposed by Fang et al [9]. CNN-based approaches use the top layer of ConvNet for extracting for extracting information of the salient objects from the image. These techniques may lose information which is useful to generate more detailed captions. This can be overcome by using features from the lower convolutional layer instead of fully connected layer [1].

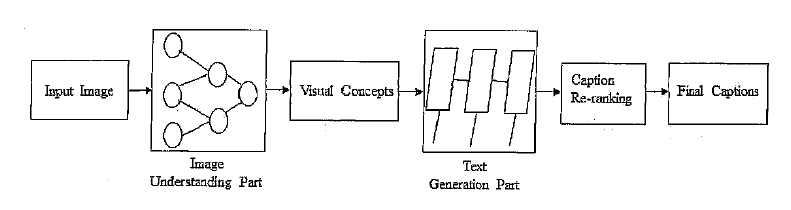


Figure 4. Generic Compositional architecture-based model [1].

An attention based mechanisms are becoming popular in deep learning image captioning. They focus on the various parts of the input image while the output sequences are being produced. A typical method follows these steps [1]: 1) Image information is obtained based on the whole scene with a CNN 2) The language generation phase generates words or phrases based on the output from 1. 3) Salient regions of the image are focused in each time step of the language generation model based on generated words or phrases. 4) Captions are generated dynamically until the end state of the language generated model.

The main difference between attention methods from others is that they can concentrate on the salient parts of the image and generate corresponding words at the same time. This uses stochastic hard attention and deterministic soft attention to generate attentions.

Anderson implements bottom up-attention using Faster R-CNN. Faster R-CNN detects objects in two stages. Faster R-CNN is used in conjunction with ResNet-101 CNN to generate an output set of image features for use in image captioning. Faster R-CNN detects objects in two stages. First, a Region Proposal Network (RPN, predicts object proposals. In the second stage, Region of Interest Pooling (RoI) is used to extract a small feature map for each box proposal. The feature maps are then boxed together are as input for final layers of the CNN.

Semantic concept based have become more popular. These methods selectively attend to a set of semantic concept proposals extracted from the image. The concepts are combined into hidden states and the outputs of recurrent neural networks. These methods follow these steps [1]: A CNN based encoder is used to encode the image features and semantic concepts. The image features are fed into the input of language generation model. Semantic concepts are added to the hidden states of the language model. Then, the language generation part produces captions with semantic concepts.

You et al [10] propose an image captioning model with semantic attention. It was novel in that it combined bottom up and top down approaches. Top-down approaches start from a gist of an image and convert into words, and bottom up come up with words describing various aspects of an image and then combine them. This model learns to selectively attend to semantic concept proposals and fuses them into hidden states and outputs of RNNS. This forms a feedback connecting the top-down and bottom-up computations. Bottom-Up and Top-Down approaches are also being used in other captioning models and for other tasks such as visual question answering (Anderson, He, Teney, et al) [11]. Top-down visual attention mechanism enable deeper image understanding through fine-grained analysis. These are what most conventional visual attention mechanisms are (Anderson). These methods are typically trained to selectively attend to the output of one or more layers of a CNN. This corresponds to a uniform grid of neural receptive fields regardless of the content of the image.

General Adversarial Networks (GAN) is a method previously used in other image tasks. These are unsupervised learning methods. The aim to improve the naturalness and diversity of captions. GANS and conditional GANS (CGANS). The first to use CGANs to generate image descriptions were Dai et al. [12]. The generator G and an evaluator E. Given an image I, the generator G takes two inputs: an image feature f(I) derived from a CNN and a random vector z. The random vector allows the generator to produce different descriptions given an image. The diversity of the cations can be controlled by altering the variance of z. The generated uses an LSTM net a decoder to generate a sentence word by word. The evaluator E is another neural network. It embeds the image I and words of the sentence into vectors of the same dimensions. The image with an CNN and the sentence with an LSTM net. The quality of the description is measured by the dot product of the embedded vectors. A general form

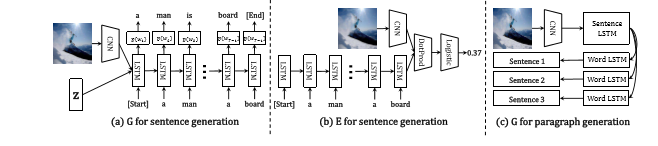


Fig 5. The structures of the generator G for both single sentences and paragraphs, and the evaluator E for single sentences.

Wang et al [13] proposed a parallel-fusion RNN-LSTM architecture for image caption generation. Combining the advantages of simple RNN and LSTM. The proposed approach divides the hidden units of RNN into several same-size parts, and lets them work in parallel. Then, the outputs are merged with corresponding ratios to generate final results. Moreover, these units can be different types of RNNs, for instance, a simple RNN and a LSTM.

Novel object caption generation methods can generation descriptions of objects which are not present in paired image-caption datasets. Generally, these methods follow the steps [1]: A separate lexical classifier and a language model are trained on unpaired image data and unpaired text data. Then, deep caption model is trained on paired image caption data. Both models are combined together to train jointly to generate captions for novel objects. This is different than many captioning methods which cannot present objects unseen in the test images.

Hendricks et al. [14] proposed one of the first models in this class. It explicitly transfers the knowledge of semantically related objects to compose the descriptions about novel objects in the proposed Deep Compositional Captioner (DCC).

Yao et al [15] proposed a new architecture which uses a Long Short-Term Memory with Copying Mechanism (LSTM-C). It incorporates copying into the Convolutional Neural Networks (CNN) plus Recurrent Neural Networks (RNN) image captioning framework, for describing novel objects in captions. Specifically, freely available object recognition datasets are leveraged to develop classifiers for novel objects.

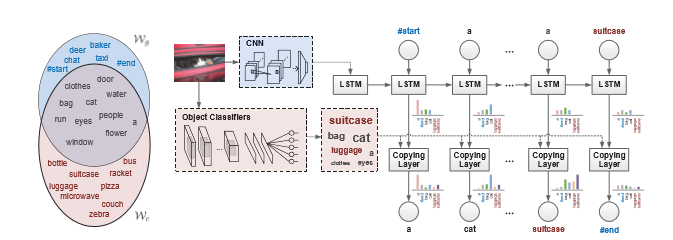


Fig 6. LSTM-C model. Wg and Wc are the vocabularies on paired image-sentence dataset and unpaired object recognition dataset, respectively. The image representation extracted by CNN is injected into LSTM at the initial time for standard word-by-word sentence generation. The object classifiers learnt on unpaired object recognition dataset are used to detect the object candidates which are additionally incorporated into LSTM, enabling captioning of novel objects. A copying mechanism and end-to-end trainable architecture leverages the standard word-to-word sentence generation mechanism as well as the newly developed copying mechanism.

One area of research with image captioning involves making captions more engaging. This is the area of stylized caption generation. A model developed by Shuster et al [16] aims to define a new task “Personality-Captions” where the goal is to be as engaging as possible to humans by incorporating controllable style and personality traits. The researchers collected a dataset of over 240 thousand captions conditioned over 215 possible traits. They developed an overall TransResNet model (ResNeXt-IG-3.5B), shown below. The researchers designed different models—generative and retrieval. The generative model gives a new state of the art on COCO caption generation, and the retrieval architecture, TransResNet, yields the highest known R@1 score in the Flickr30k dataset. Another method that have worked to make captions more engaging have included using puns [17]

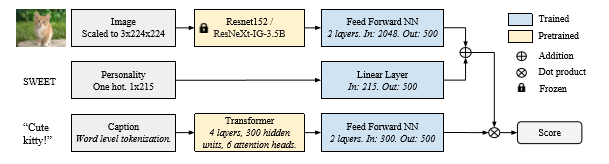


Figure 7: TransResNet architecture, used for our retrieval models.

* 1. Evaluation Metrics

BLEU- The BLEU (Bilingual evaluation understudy) Score, originally described in 2002, is a common method for automatic evaluation of machine translation [18]. Individual text segments are compared with a set of reference tests which are human generated and scores are computed; the overall score is the average of the scores computed. The score should be between 0 and 1, one being a perfect match. Bleu is a popular measure but is better used for short texts [19]. BLEU scores are generally reported as a set of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. BLEU computes the geometric mean of n-grams precision (1 ” n ” N) with a positive weighting.

ROUGE- (Recall-Oriented Understudy for Gisting Evaluation) [20]- This metric measures the quality of a text comparing word sequences, word pairs, and n-grams with a set of reference summaries created by humans. ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S worked well in single document summarization tasks, while ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, and ROUGE-SU9 work better for evaluating very short summaries. Multi-document evaluation is not well evaluated with ROUGE.

METEOR [21]- METEOR was designed to address weaknesses identified with BLEU. Namely, lack of recall, using higher order n-grams as an indirect measure of how well a translation is grammatically formed, Lack of Explicit Word-matching Between Translation and Reference, and the use of Geometric Averaging of N-grams. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported

SPICE (Semantic Propositional Image Captioning Evaluation) [22] - is a newer evaluation metric, developed in 2016. It is based on a graph-based semantic representation called scene-graph (Anderson) the newly proposed SPICE correlates well with human judgments, but fails to capture the syntactic structure of a sentence [23].

CIDEr (Consensus Based Image Caption Evaluation) [24]- This method introduces a novel consensus-based evaluation protocol, which measures the similarity of a sentence to the majority, or consensus of how most people describe an image.

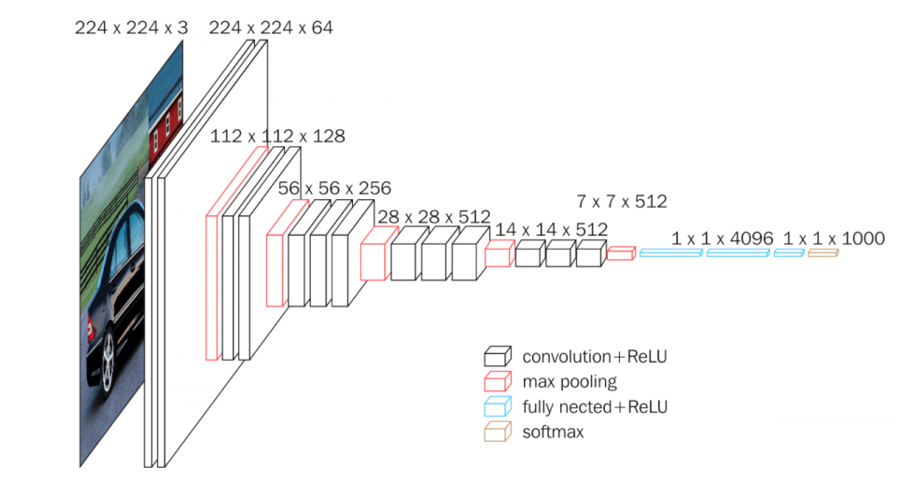
Metric developed by Cui et al [23]- Cui et al proposed a novel learning-based discriminative evaluation metric directly trained to distinguish between human and machine-generated captions. It correlated with human judgements while also capturing the syntactic structure of a sentence.

3. Methods

Here a CNN-LSTM model is developed to generate novel captions from images. A VGG16 model trained on the ImageNet dataset is used. The input to Convolution1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional layers, where the filters are used with a very small receptive field: 3×3. This is the smallest size to capture the notion of left/right, up/down, center. The convolution stride is fixed to 1 pixel; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 convolutional layers. Spatial pooling is carried out by five max-pooling layers. Max-pooling is performed over a 2×2 pixel window, with stride 2.

Three Fully-Connected (FC) layers follow a stack of convolutional layers. The first two have 4096 channels each, the third performs 1000-way ILSVRC classification and contains 1000 channels (one for each class). The final layer is the soft-max layer. All hidden layers are equipped with the rectification (ReLU) non-linearity.

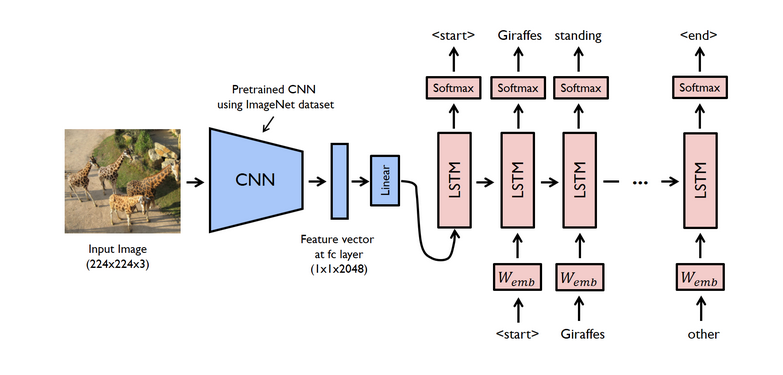
The VGG16 model is visualized below.



**Fig. 9 [25]** VGG16 architecture



**Fig. 10 [25]** Illustration of ConvNet configuration with pooling



**Figure . Generalized model CNN-LSTM used here. The CNN here is VGG16.**

The last layer is removed from the loaded model is removed to get the features. Image size is converted to 224\*224, image pixels are converted to a numpy array and the image data is reshaped for the model. A preprocess function is used to convert the images to the format required. Features are stored in a dictionary. The extracted features from the photo feature extractor are used as input.

To prepare the text data, the descriptions are loaded into memory a mapping of photos to descriptions is created. Descriptions are cleaned for example by removing converting to all lower case, removing one letter words, and removing words with a number in them which may be misspelled. The loaded descriptions are converted into a vocabulary of words. The sequence processor is a word embedding layer for handling text input, followed by a LSTM recurrent neural network layer.

The feature extractor and sequence processor both output a fixed-length vector. These are merged together and processed by a dense layer to make a final prediction. The Photo Feature Extractor model expects input photo features to be a vector of 4,096 elements. These are processed by a dense layer to produce a 256 element representation of the photo.

1. Results and Discussion



**example of a good caption:**

**dog is running in the water**

**example of a poor caption: man in red shirt is riding bike on the street**

**BLEU-1: 0.539220**

**BLEU-2: 0.284752**

**BLEU-3: 0.190485**

**BLEU-4: 0.085424**

**Figure . Sample images with generated captions.**

Best

Due to its depth and number of fully-connected nodes, VGG16 is over 533MB. This makes deploying VGG a tiresome task.VGG16 is used in many deep learning image classification problems; however, smaller network architectures are often more desirable (such as SqueezeNet, GoogLeNet, etc.)

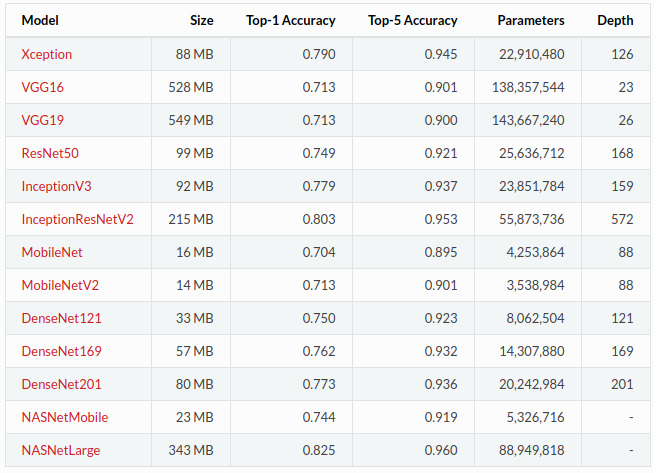
The VGG model It has so many weight parameters, the models are very heavy, 550 MB + of weight which also means long inference time. Resnets take lesser memory, faster inference time, and allows deeper networks to be trained. Based on your problem, you can decide how many layers you want for the required accuracy and inference time requirements. Its rather simple to understand too.

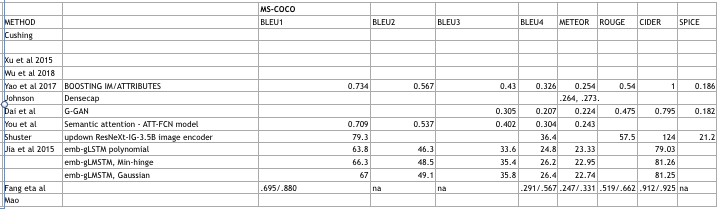
Several state-of-the art models are shown with their metric scores. Here an encoder-decoder, which is a common type of model, is implemented. It is questionable whether these types of models can deal with the two types of modalities—the structure of the visual information is very different from the structure of the description to be generated (Jia et al). Jia et al note that CNN-LSTM encoder-decoders sometimes generate sentences that seemed to “drift away” or “lose track” of the original image content generated a description that is common in the dataset but only weakly coupled to the input image.

Jia et al use alter the LSTM model by adding semantic information as an extra input to the gate of each LSTM memory cell. This input can take different forms as long as they build a semantic connection between the image and its description, a semantic embedding, a classification or a retrieval result. Jia as observes that current methodologies are heavily biased towards short sentences and this hurts the quality, thus they propose sentence normalization. The three different normalization methods found most useful are report in their results: polynomial, min-hinge, and Gaussian. These normalization methods when couples with a gLSTM brings much improvement in the performance of the captioner. Their model performs on par with more complicated and expensive attention mechanisms and other methods which use an emsemble of multiple LSTM models (Vinyals et al 2015).

Attention mechanisms are designed to align the information in both the source and target domains, so that the model is able to attend to the most relevant part of the sentence from source language or image.

Shuster’s work, “Engaging Image Captioning via Personality” demonstrates one method in which captions can be made more engaging. Shuster’s used current state of the art models both for retrieval models and generative models. Retrieval models produce a caption by considering any caption in the training set as a possible candidate, while generate models generate word-by-word sentences conditioned on the image and personality trait, using a beam. Both require a caption encoder, a Transformer architecture is employed here. For the generative model evaluated on the new personality captions dataset, all scores are lower for the personality-captions dataset (not shown) than on the COCO dataset.



Figure

**5. Conclusion and Future Scope**

**REFERENCES**

[1] Z. Hossain, MD, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Comput. Surv*, vol. 51, no. 6, Feb 2019, 36 pages, art no. 18.

[2] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. InEuropean conference on computer vision 2014 Sep 6 (pp. 740-755). Springer, Cham.

[3] J. Deng, W. Dong, R. Socher, L. Li ,K. Li ,and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database” In *2009 IEEE conference on computer vision and pattern recognition,* Jun 2009, pp. 248-255,. IEEE.

[4] R Krishna, et al, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International Journal of Computer Vision*, vol. 123, no. 1, May 2017, pp. 32-73.

[5] P. Sharma, N. Ding, S. Goodman, and R Soricut. “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning”. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,*  Vol. 1: Long Papers, Jul 2018, pp. 2556-2565.

[6] J. Mao, W. Xu, Y. Yang, J Wang , Z Huang, and A.Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn)”, arXiv preprint arXiv:1412.6632. 2014 Dec 20.

[7] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* *2016*, pp. 4565-4574.

[8] X. Jia, E. Gavves, B. Fernando, T. Tuytelaars, “Guiding the long-short term memory model for image caption generation,” In *Proceedings of the IEEE international conference on computer vision* *201,* pp. 2407-2415.

[9] H Fang, et al, “From captions to visual concepts and back” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473-1482.

[10] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016, pp. 4651-4659.

***[11]Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang***. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077-6086.

[12] Dai Z, Almahairi A, Bachman P, Hovy E, Courville A. Calibrating energy-based generative adversarial networks. arXiv preprint arXiv:1702.01691. 2017 Feb 6.

[13] M. Wang, L. Song, X. Yang, and C. Lu, “A parallel-fusion RNN-LSTM architecture for image caption generation,” In *2016 IEEE International Conference on Image Processing (ICIP)*, Sep 2016, pp. 4448-4452. ieee

[14] Anne Hendricks L, Venugopalan S, Rohrbach M, Mooney R, Saenko K, Darrell T. Deep compositional captioning: Describing novel object categories without paired training data. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 1-10).

[15] Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. InProceedings of the IEEE

International Conference on Computer Vision 2017 (pp. 4894-4902).

[16] *K. Shuster, S. Humeau, H. Hu, A. Bordes, Jason Weston*. Engaging Image Captioning via Personality. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12516-12526.

[17] A. Chandrasekaran A, D. Parikh, M. Bansal, “Punny captions: Witty wordplay in image descriptions,” arXiv preprint arXiv:1704.08224. 2017 Apr 26.

[18] K Papineni, S Roukos, T Ward, and W Zhu, “BLEU: a method for automatic evaluation of machine translation,” In *Proceedings of the 40th annual meeting on association for computational linguistics, Jul* 2002 pp. 311-318, Association for Computational Linguistics.

[19] C. Callison-Burch, M. Osborne, and P. Koehn, “Re-evaluation the Role of Bleu in Machine Translation Research,” In *EACL*, Vol. 6, pp. 249–256.

[20] C. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8. Barcelona, Spain.

[21] S. Banerjee, A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. InProceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization 2005 Jun (pp. 65-72).

[22] Anderson P, Fernando B, Johnson M, Gould S. Spice: Semantic propositional image caption evaluation. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 382-398). Springer, Cham.

[23] ***Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie***, “Learning to Evaluate Image Captioning,” *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5804-5812.

[24] R. Vedantam, L. Zitnick C, D. Parikh. “Cider: Consensus-based image description evaluation,” In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2015, pp. 4566-4575.

[25] <https://neurohive.io/en/popular-networks/resnet/> Resnet (Residual CNNs for Image Classification Tasks)

[23] Stack Exchange: Data Science. Which is the fastest image pretrained model? October 2018. Retrieved from. https://datascience.stackexchange.com/questions/39177/which-is-the-fastest-image-pretrained-model

Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures

by Bernardi, Raffaella; Cakici, Ruket; Elliott, Desmond; More...

Journal of artificial intelligence research, 2016, Volume 55

D. Liu, Z. Zha, H. Zhang, Y. Zhang, and F. Wu. “Context-Aware Visual Policy Network for Sequence-Level Image Captioning”, in *Proceedings of the 26th ACM international conference on Multimedia* (MM '18). ACM, New York, NY, USA, 1416-1424. DOI: <https://doi.org/10.1145/3240508.3240632> 2018

Elliott, Desmond & Keller, Frank. (2014). Comparing Automatic Evaluation Measures for Image Description. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 2. 452-457. 10.3115/v1/P14-2074.

Staniūtė, Raimonda; Šešok, Dmitrij. A Systematic Literature Review on Image Captioning. *Applied Sciences*, 05/2019, Volume 9, Issue 10.

Guan, Z.; Liu, K.; Ma, Y.; Qian, X.; Ji, T. Middle-Level Attribute-Based Language Retouching for Image Caption Generation. Appl. Sci. 2018, 8(10), 1850; https://doi.org/10.3390/app8101850.

https://www.mdpi.com/2076-3417/8/10/1850.

Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Nezami, Omid Mohamad, et al. "Face-Cap: Image Captioning Using Facial Expression Analysis." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2018.

Anderson, Peter, Stephen Gould, and Mark Johnson. "Partially-supervised image captioning." *Advances in Neural Information Processing Systems*. 2018.

Gu, J., et al. "Stack-captioning: Coarse-to-fine learning for image captioning." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Priya Dwivedi. Jan 4. Understanding and Coding a ResNet in Keras. <https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33>

Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In Advances in Neural Information Processing Systems, pages 2361–2369, 2016.

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evalu- ation metrics. Journal of Artificial Intelligence Research, pp. 853–899, 2013.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78, 2014.

Gu J, Wang G, Cai J, Chen T. An empirical study of language cnn for image captioning. InProceedings of the IEEE International Conference on Computer Vision 2017 (pp. 1222-1231).

O. Vinyals, A Toshev, S. Bengio, and D. Erhan

VGG-Convolution Network for <https://neurohive.io/en/popular-networks/vgg16/>

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

Appendix I