A Survey on Different Deep Learning Architectures for Image Captioning

NIVEDITA M., ASNATH VICTY PHAMILA Y.
Vellore Institute of Technology, Chennai, 600127, INDIA

Abstract: Vision plays an important part which helps us to look at the world and perceive information about our surroundings. A human perceives information by looking at an object or the surrounding on the whole and tries to map visual features and attributes and by summarizing these features we can describe or tell about our surroundings. The way the human brain does this is still a huge mystery. But, For a machine/computer this task is what is called as Image Captioning. The computer or machine is fed with images from which they learn to extract features i.e pixel information, object position, geometry, etc. Using these features the machine tries to map it to a sentence word by word or on a whole which summarizes the information of the image. Due to the advancements in recent Computer Vision Methods and Deep Learning architectures, Computers have been able to correctly summarize images which have been fed to them. In this paper, we present a survey on the new types of architectures and the datasets which are being used to train such architectures. Furthermore, we have discussed future methods that can be implemented.

Keywords: Computer Vision, Image Captioning, Convolutional Neural network, Recurrent Neural network

Received: May 11, 2020. Revised: October 13, 2020, Accepted: November 6, 2020. Published: November 19, 2020.

1. Introduction

Image Captioning is the process of perceiving various relationships among objects in an Image and give a brief description or summary of the image. When a person is asked to describe an image, he/she first notices the various objects in the given image. Next they analyze how the objects are positioned, their physical appearance and how each object in the image is related to the other objects based on position, geometry, etc. But a computer cannot carry out all these tasks. This is where Image Captioning comes into place, where Deep Learning methods are used to analyze the images and come out with textual descriptions for each image. These Deep Learning models follow a standard structure with few modifications. The whole model usually consists of two sub-models: A Encoder (CNN) for extracting features from the image, A Decoder (NLP Language Model) for generating the captions based on the input features. The output of the Encoder is directly passed to the NLP Language Model along with the train captions during training. The research extent in Image Captioning has increased tremendously and a number of models with a variety of methods have been proposed. All these models follow the Encoder (CNN) - Decoder (Language Model) architecture. The naive way is directly feeding the output features from the Encoder Model into the Language Model along with the captions. Along with this model architecture, attention models are also implemented to imitate the visual attention mechanism of a real person to capture salient features and attributes when generating a word based on the image. The models proposed till now follow a CNN + RNN or CNN + LSTM architecture setup. Another model architecture was proposed in 2018 which made use of CNN - CNN architecture where the Encoder CNN was used to extract features from images and the Decoder CNN (Language CNN) was used

in place of traditional RNNs or LSTMs. In 2017 a paper titled ”Attention Is All You Need” [27] introduced a novel architecture called the Transformer. Since then studies have proven that Transformer Architectures perform much better than RNNs or LSTMs.
Thus the main contributions of this paper are as follows:

- Introduction and Overview to the different types of Architectures used till now
- An Overview to the Transformer Architecture
- A Literature Survey on the different Transformer Models in Image Captioning
- Summarized the Datasets and Evaluation Metrics being used
- A discussion on future improvements

2. Literature Survey

Since the year 2014, where the first Deep Learning based Image captioning architecture was proposed by Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel [10], Many different types of architectures have been proposed by researchers. These architecture categories include Encoder - Decoder architectures, Attention Based architectures and Multi - Modal architectures. In 2017, a new type of architecture called the Transformer architecture emerged and since then Transformers have proved to best results of typical RNNs and LSTM based systems. In this section we shall summarize all these architecture types and give a overview on some of the models based of these architectures.

*2.1. **Encoder - Decoder Architecture***

This architecture is divided into two parts. One is the Encoder (Convolutional Neural Network) and the other is the Decoder (Language Model such as RNNs, LSTMs). The Encoder or the CNN takes the images as input and outputs image features as a vector feature map which is the result of the hidden activations of the CNN. The Encoder can also be used to output relationship attributes of the objects in the images. These features along with their relationships and the corresponding image captions are passed as input to a Decoder Model which is a RNN or a LSTM language model. The Decoder trains on this input data and tries to predict each word of the image caption at each timestep. This type of model is represented in Fig. 1

$$\mathbf{x} = \mathit{Encoder}(\mathbf{img})$$

$$\mathrm{x}_{\mathrm{t}} = \mathbf{WS}$$

$$\mathbf{P} = \mathit{Decoder}(\mathrm{x}_{\mathrm{t}})$$

vocabsize - the vocabulary size, Where **img** - the input image, **x** - feature map which is passed as input to the Decoder, **S** - one-hot vector with **vocabsize** which represents the **t**-th word of the image description and **S0** -the [START] tag and **Sn**-the [END] tag. **W** - the embedding matrix ; and **P** matrix represents the probability vector that is generated at t+1 time step .

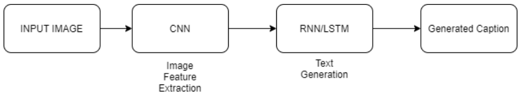

Figure 1: A standard image captioning architecture

We will have an overview of few of the models based on this architecture:

- **AlexNet - MLBL** : In 2014, Kiros et al. [10] proposed a model which makes use of the AlexNet [12] as the Encoder CNN to extract image features as a feature map. They then pass this feature map along with the word representation vectors of the sentence to their proposed Log Bilinear Model (LBL).This was the first model which was proposed in the field of image captioning and is a deterministic one. It is a feed-forward neural network and it has only one linear hidden layer. The LBL operates on word representation vectors similar to other neural language models. Further adding to this, they came up with two more models under the collecitve name Multimodal Log-Bilinear Models. The two models are as follows:
  - **Modality-Biased Log-Bilinear Model (MLBL-B)** In The MLBL-B the feedforward

network of the LBL is taken as such and then an additional context channel is added based on the modality x.

- **The Factored 3-way Log-Bilinear Model (MLBL-F)** This model incorporated modality conditioning by gating the word representation matrix R by the features x.

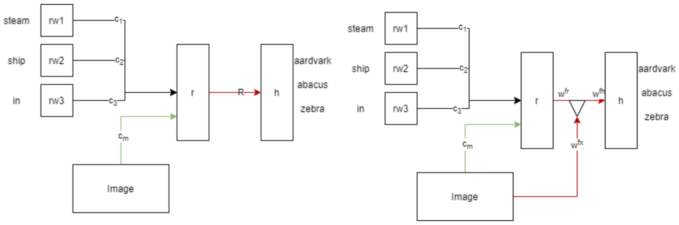

Figure 2: Left : MLBL-B Model and Right: MLBL-F Model as given in [10]

- **VGGNet - RNN** :After the first model came out In 2014, Karpathy et al. [9] proposed a Model that aligns the sentence or caption snippet corresponding to the visual regions that they represent through multimodal embeddings. This model uses a RCNN (Region Convolutional Neural Network) (VGGNet) [24] pretrained on the ImageNet dataset to detect objects in every image.This produces vector with 4096 dimensional activations which is taken as the input features to the Decoder Model. To ensure that the Decoder Model establishes inter-modal relationships they make use of BRNN (Bidirectional Recurrent Neural Network). A sequence of N words which is encoded as 1-of-k representation is taken as input by the BRNN and transforms it into a n-dimensional vector. Then the model tries to map each feature of the image to the a word in the sentence and computes a measure of similarity and takes this score and computes overall score between the given image and sentence.

- **ResNet - LSTM** : Jiasen et al. in their paper [20] had used a attention based ResNet [6] - LSTM [7] Model which automatically decides when to rely on visual signals and when to just rely on the language. Along with this encoder - decoder architecture they have proposed a spacial attention mechanism for extracting spatial image features. This attention model along with the LSTM produces a ”visual sentinel” vector instead of just the hidden activation states. It acts as a fallback option for the decoder. model.

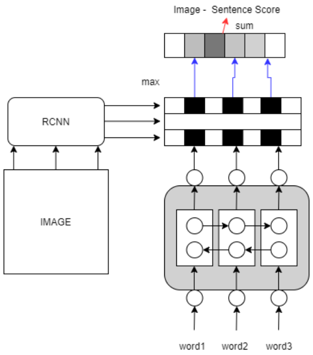

Figure 3: CNN - BRNN Model Architecure used in [9]

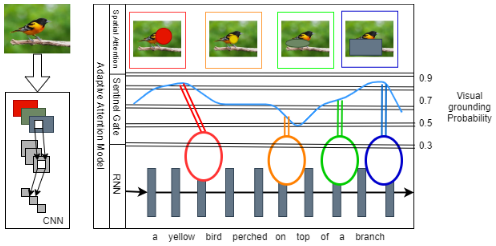

Figure 4: The Model Architecture with Spatial attention mechanism for finding when and where to look for word generation. [20]

- **InceptionV3 - LSTM** : In the paper [18], Liu et al. propose a new evaluation metric called the *SPIDEr*. This score is a combination of the SPICE and CIDEr metrics. This score provides a evaluation metric that 1) assigns a caption with high score if it is considered good by humans, 2) And the captions with high scores are considered good by humans. And secondly they propose a Policy Gradient that can optimize the captioning metrics. The Encoder - Decoder Architecture used here is a InceptionV3 [25] - LSTM Decoder Architecture and similar to that of Show and Tell architecture [29] . The CNN is pretrained on ImageNet and the LSTM uses about 512 units.

- **CNN - CNN** : In the paper [30], Wang et al. proposed a CNNs (Convolutional Neural Networks) only encoder - decoder. They had used a CNN as the decoder also since RNNs or LSTMs cannot be calculated in parallel and ignore the underlying hierarchical structure of a sentence.

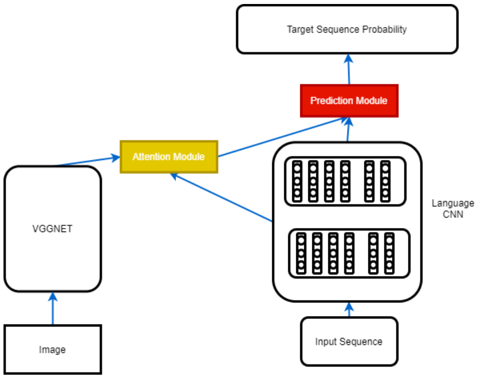

Figure 5: CNN+CNN model for image captioning. The vision CNN extracts features from the image, and the language CNN is applied to model the sentence.

2.2. *Attention Based Architecture*

In the year 2015, Bahdanau proposed an architecture [2] called the encoder-decoder model which predicts the target word by automatically searching all the parts of a sentence whatever is relevant to the target word. This is termed as Attention where the model pays attention to relevant words in the sentence.

A Bidirectional LSTM is used to generate a sequence of annotations *(h1,h2,...hTx)* for each input sentences. These vectors i.e *h1,h2,...hTx* are basically the representations of **Tx** number of words in the sentence which is the result of concatenating both encoder's forward and backward hidden states. Then the importance score for each candidate vector was calculated and then the scores are normalized to weights using the softmax function.Then the final step is that these weights are applied to the candidates to generate the attention result which is a weighted average vector

We will have an overview of few of the models based on this architecture:

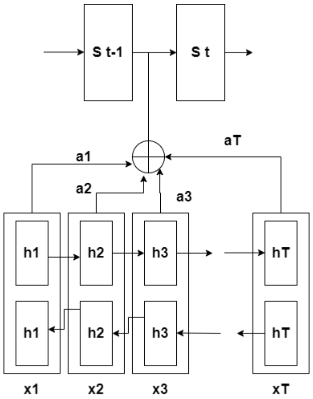

Figure 6: Attention Model as given in Bahdanau's paper [2]

- **Text Guided Attention Model** : In 2016, Jonghwan et al. in their paper "Text-guided Attention Model for Image Captioning" [21] proposed a novel attention model called the *text-guided attention model* which exploits the captions in the training set which acts as a source for visual attention. It is a sampling-based scheme to learn attention using multiple the guidance captions taken from the training set. This prevents overfitting in training and removes the problem of learning unwanted attention from noisy captions.

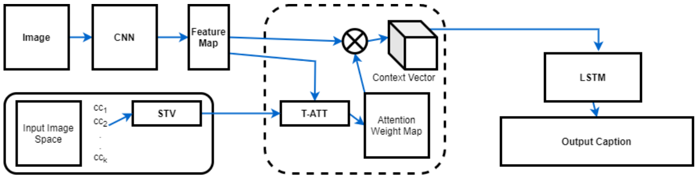

Figure 7: Text Guided Attention model as given in [21] The text-guided attention layer (T-ATT) computes an attention weight map. A context vector is obtained by aggregating image feature vectors weighted by the attention map. Using this context vector the LSTM decoder generates an output caption.

- **Semantic Attention Model** : Quanzeng et al. in their paper "Imgae Captioning with Semantic Attention" [32], proposed an approach which combines both the top-down as well as bottom-up approaches through a semantic attention model. This provides a detailed description of objects which apper semantically important which are to be found exactly where they are needed. The main focus of this attention model was to attend to semantically important regions, ability to weight the relative strength of the attention on objects and the ability to switch attention among objects dynamically. The information from the top-down feature will be captured by the initial state which is located on the top of Recurrent Neural Network (RNN). From the bottom up attributes the RNN states will get the feedback and interaction through an attention mechanism that is enforced on both network state and output nodes. This feedback from the RNN helps the algorithm in predicting the new words more accurately and also providing more robust inference of the semantic gap that is existing between already existing predictions and image content.

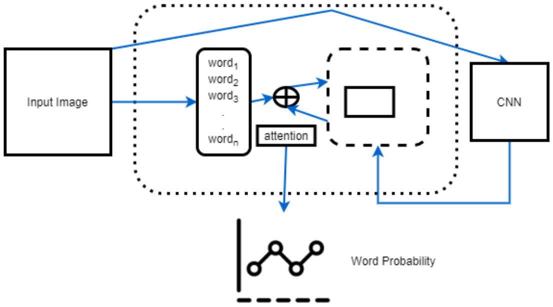

Figure 8: Semantic Attention model as given in [32]. The visual features are combined wit visual concepts by Semantic Attention model and the recurrent neural network generates the image caption from that.

- **Explicit Attention Model** : In 2017, Chenxi et al. proposed a ”explicit attention model” [17] that can not only be used on detailed ground truth attention maps such as the Flickr30k dataset [33], but also when only the semantic labelings of image i.e MSCOCO dataset [16]. This model was proposed to tackle the problem in the model proposed in [31] that is we cannot able to learn a better attention regions even if we have some prior knowledge about the attention map.

2.3. *Transformer Architecture*

In 2017 Vaswani et.al [27] in their paper titled ”Attention Is All You Need” introduced a novel architecture called the Transformer. As per the title describes, the paper makes use of attention mechanisms. But unlike the previous models this model makes use of two parts encoder and decoder to transform one sequence into another one and does not have any Recurrent Neural Networks.

As we can see in 9 the Encoder is on the left and decoder on the right. Both these components have modules which are stacked on top of each other. These modules are either Multi - Head Attention modules or Feed Forward layers.

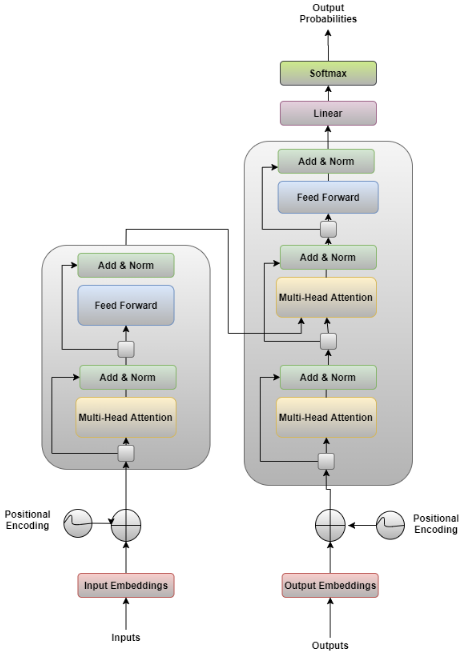

Figure 9: Transformer Model Architecture as given in [27]

The SoftMax function is usually applied to the weights a to have a distribution between 0 and 1. The resultant weights from the softmax function are then applied to all the words in the sequence.

A pointwise feed-forward layer is introduced after the multi-attention heads in both the encoder and decoder. This little feed-forward network has identical parameters for each position, which can be seen as a separate, identical linear transformation of each element from the given sequence.

We will have an overview of few of the models based on this architecture:

- **Transform and Tell**: One of the recent papers of 2020, ”Transform and Tell” by Alasdair et al. [26] introduces the use of a special transformer architecture as their Text Encoder and a ResNet-152 as their Image Encoder and trained this model on a news article dataset known as the **NYTimes800k**. Along with this they use MTCNN as their Face Encoder to detect faces, Since news articles also focus on celebrities and people. This is done so that the model can pay attention and incorporate the relationships of named entity texts while generating the news captions/descriptions.
  The transformer used in this paper is known as **RoBERTa** [19] which is a pretrained language representation model which provides contextual embeddings for text. This is a recently improved version of the **BERT** [5] Model. It consists of 24 layers of bidirectional transformer blocks. This bidirectionality along with the attention provided by the multi head attention blocks in the transformer, allow a word to have multiple vector represetations depending on the context of it’s surroundings.

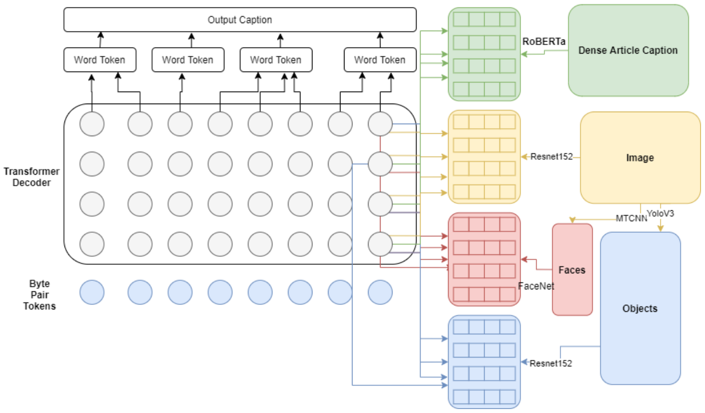

Figure 10: Overview of the Transform and Tell Model [26].

- **Boosted Transformer**: In 2019 Jiangyun et al. in their paper ”Boosted Transformer for Image Captioning ” [14] came up with a modified version of the transformer architecture. On the whole it is a transformer based encoder-decoder architecture. A Faster R-CNN is used first to detect a set of image region and then it obtain the aligned visual features as well as the semantic concepts from the image . The term ”Boosted” is because they use a concept-guided attention module which has two self-attention mechanisms and an integration module . The result from these model will be the boosted visual features. After this the decoder module uses these boosted features and the sequence representations as input to a sequence of attention modules and a feed-forward network in order to generate a caption.

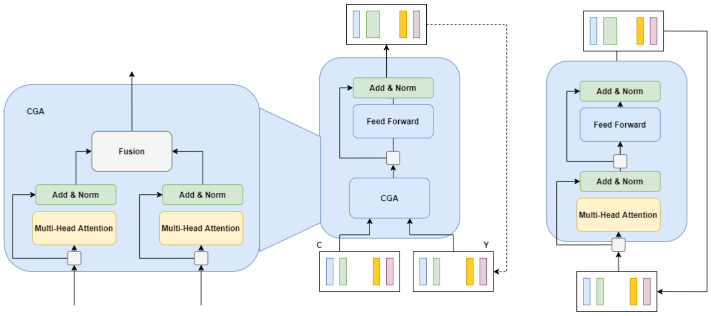

Figure 11: The Boosted Transformer Model [14].

- **Entangled Transformer**: Li et al. had proposed an Entangled Transformer Model [13] in 2019. This model makes use of a dual-way encoder as the image encoder. It consists of two sub-encoder with each being self - attentive. Another one is the decoder block which has the self-attention sub-layer and the feed-forward sub-layer and an ETA module and a GBC module between them. This feature enables the decoder block to carry out attention over the image encoder's visual and semantic outputs. An overview of the model can be seen in Fig. 12

- While paying attention to the target one,the Entangled Attention Module implements the attention in an entangled manner so that it can be affected by the preliminary modality.

- Gated Bilateral Controller (GBC) specially designed for the integration of the generated representations st and vt.This module is similar to the gates that are present in LSTMs, GRUs, etc. These gates are efficient in dealing with vanishing and explosion gradient. Thus these gates enables the propagation of the information through long timesteps or many deeper layers.

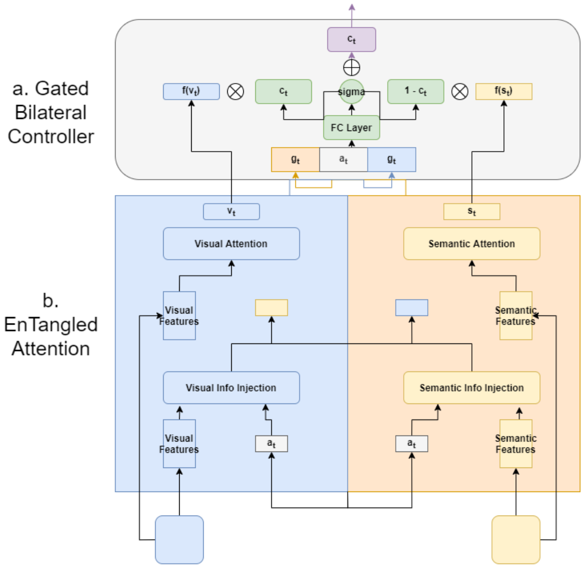

Figure 12: ETA (b)to conduct EnTangled Attention, then to GBC (a) to obtain the final representation.

- **Meshed-Memory Transformer**: Marcella et al. had proposed a Meshed Memory Architecture [4] where encoder is in charge of processing The regions from the input image will be processed by the encoder and it devises the relationships between the regions and the output of each encoding layer will read by the decoder and the output caption will be generated word by word. The Encoder - Decoder used here is as follows:

  - **Encoder with Memory Augmented Attention** : In this model,a set of image regions X is extracted from an input image and attention is applied to them inorder to obtain a encoding of X through the self-attention operations that is used in the Transformer.some additional "slots" are used along withe the set of keys and values which is used for self-attention and can encode a priori information. To ensure that a priori information should not depend on the input set X, The additional keys and values used are implemented as plain vectorsin order to ensure that the priori information should not depend on X,the input set. The vectors can be directly updated via Stocastic Gradient Descent.

  - **Decoder with Meshed Cross-Attention** : The Meshed Attention operator in the decoder side connects Y, the input vector sequence to all elements in X,output from all encoding layers through gated cross-attention.

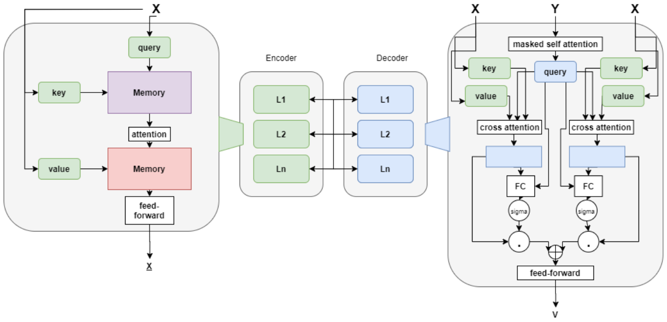

Figure 13: The Meshed Memory Architecture as give in [4]

3. Datasets

Image captioning requires large amounts of data i.e Image - Caption pairs which are passed as inputs to the different model architectures for training. These datasets are made available by researchers and organizations who have tagged each image with a set of descriptive captions. The datasets that are available are: Flickr8K, Flickr30K, MS COCO, Visual Genome and Google Conceptual Captions.

Flickr8K[8]
It contains 8092 images as a total collected from official site Flickr.com. Each image in the dataset have 5 captions associated with it which were obtained through Amazon Mechanical Turk services. The average length of the sentences is about 11.8, which each caption having a accurate description of the objects and the scene depicted in the image. Thus a total of 40,460 captions are present in the dataset. In practical implementations the whole dataset is divided into Training Set with 6000 images, Dev Set with 1000 images and Testing set with 1000 images.

Flickr30K[33]
This dataset is an extended version of the Flickr8K and contains a total of 31783 images collected from Flickr.com (including Flickr8K images) with each image having 5 captions thus a total of 158,915 descriptions . In practical implementations Dev Set uses 1000 images and Testing set uses 1000 images. The rest are used as training set. This is because the rule of training models with large datasets is that use a small fraction for testing and dev sets and the rest of them are used for training.

Microsoft COCO[16]
Microsoft COCO (Common Object in COntext) is the dataset released by the Microsoft research team. This is one of the most popular datasets which is being used for many tasks such as object detection, image segmentation,instance segmentation and image captioning. This dataset consists about 91 categories with a total of 328K images, 2.5 million tag instances and 5 captions are associated with each image.Released in 2014, it consists of 82,783 train set images and 40,504 as validation set images and 40,775 as test set images. Since the test set doesn't have descriptions it the train and validation set is further divided into train/validation/test sets.

Visual Genome[11]
The size of the dataset is more than 108K images. There will be an average of 35 objects associated with each image and has dense description or captions, 21 interactions between objects and 26 attributes. This dataset is used to pre-train image captioning models that introduce semantic and spatial relationships among objects.

Google Conceptual Captions[23]
Google's Conceptual Captions dataset has more than 3 million images, here also all image is paired with natural-language descriptions/captions. The style of the dataset is similar to that of the MSCOCO dataset. This dataset consists of about 3,318,333 train set images, 15,480 validation set images and about 12,559 images as the testing set which is hidden.

These datasets are used mainly for image captioning purposes. Many other tasks such as image segmentation, Visual Q-A systems, etc. can be built using these datasets.

4. Evaluation Metrics

*4.1. **BLEU**[22]*

The Bilingual Evaluation Understudy Score or the BLEU score is the most commonly used metric of evaluation for image captioning tasks. This evaluation score works by counting the matching n-grams in the candidate or predicted translation to the n-grams in the reference sentences. A 1-gram equals to one word whereas a 2-gram or b-gram refers to a word pair. Each BLEU-n score is calculated to ensure that the word or word-pair occurences in the translation match their respective references.

$$\mathbf{BLEU} = min(1, \frac{output - length}{reference - length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

*4.2. **METEOR**[3]*

Metric for Evaluation of Translation with Explicit Ordering or the METEOR score is another most commonly used metric of evaluation for image captioning. This score is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.It includes methods such as stemming and synonymy matching as well as standard exact word matching. This score solves the shortcomings of the BELU score and has a better impact on evaluating sentences.

Unigram Precision

$$\mathbf{P} = \frac{m}{\mathrm{w_t}}$$

Unigram Recall

$$\mathbf{R} = \frac{m}{\mathrm{w_r}}$$

Harmonic Mean

$$\mathrm{F_{mean}} = \frac{10PR}{R + 9P}$$

Penalty

$$\mathbf{p} = 0.5 * (\frac{c}{\mathrm{U_m}})^3$$

Final Meteor Score

$$\mathbf{M} = \mathrm{F_{mean}}(1 - p)$$

*4.3. **CIDEr[28]***

Consensus-based Image Description Evaluation or the CIDEr Score is another metric of evaluation for image captioning. This metric is a Consensus-based evaluation metric.The sentences are expressed as TF-IDF vector and the weight associated with these vectors are calculated for each n-gram .Atlast for evaluation the cosine similarity between the test set sentences and the references will be calculated.

$$CIDEr(\mathrm{c_i}, \mathrm{S_i}) = \frac{1}{m} \sum_j \frac{g^n(\mathrm{c_i}) \cdot g^n(\mathrm{S_{ij}})}{||g^n(\mathrm{c_i})||||g^n(\mathrm{S_{ij}})||}$$

*4.4. **SPICE[1]***

Semantic Propositional Image Caption Evaluation or the SPICE score is specially used for image captioning. This evaluation metric maps the objects, semantic realtionships and attributes to a graph like structure. Tuples of object/relation pairs are taken and the scores such as precision and recall are calculated. Finally using these two scores the SPICE score is calculated.

Precision

$$\mathbf{P(c,S)} = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

Recall

$$\mathbf{R(c,S)} = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

SPICE Score

$$\mathbf{SPICE(c,S)} = \mathrm{F_1}(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

*4.5. **ROUGE[15]***

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is one among a set of metrics which is used to evaluate automatic text summarization and machine translation. It will compare the automatically produced summary or translated sentence against the already existing set of reference sentences/summaries.

- ROUGE-N – unigram, bigram, trigram and higher order n-gram overlap will be measured by this metric.
- ROUGE-L – longest matching word sequences will be measured using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Predefined n-gram length is not needed because the longest in-sequence common n-grams is automatically included in it.
- ROUGE-S – It checks the pair of words in a sentence in order and it allows arbitrary gaps. This is also known as skip-gram coocurrence.

These are the most frequently used evalution metrics for machine translation tasks.

Paper Title	**Dataset Used**	**B1**	**B2**	**B3**	**B4**	**MT**	**CD**	**RG**	**SP**
Deep visual-semantic alignments for generating image descriptions.	Flickr8K	57.9	38.3	24.5	16.0	-	-	-	-
	Flickr30K	57.3	36.9	24.0	15.7	-	-	-	-
	MSCOCO	62.5	45.0	32.1	23.0	-	-	-	-
Adaptive attention via A visual sentinel for image captioning.	Flickr30K	0.677	0.494	0.354	0.251	0.204	0.531	-	-
	MSCOCO	0.742	0.580	0.439	0.332	0.266	1.085	-	-
Optimization of image description metrics using policy gradient methods.	MSCOCO	0.743	0.578	0.433	0.322	0.251	1.000	0.251	-
CNN+CNN: convolutional decoders for image captioning.	Flickr30K	0.607	0.425	0.292	0.199	0.191	0.395	0.442	-
	MSCOCO	0.685	0.511	0.369	0.267	0.234	0.844	0.510	-
Text-guided attention model for image captioning.	MSCOCO	0.749	0.581	0.437	0.326	0.257	1.024	-	-
Image captioning with semantic attention. GT-ATT	Flickr30K	0.824	0.679	0.534	0.412	0.269	0.949	0.588	-
	MSCOCO	0.910	0.786	0.654	0.534	0.341	1.685	0.667	-
Attention correctness in neural image captioning.	Flickr30K	-	-	30.2	21.0	19.21	-	-	-
	MSCOCO	-	-	37.2	27.6	24.78	-	-	-
Transform and Tell Entity-Aware News Image Captioning	GoodNews	-	-	-	6.05	-	53.8	21.4	-
	NYTimes800k	-	-	-	6.30	-	54.4	21.7	-
Boosted transformer for image captioning.	MSCOCO	81.0	65.9	51.5	39.5	29.3	130.9	58.9	23.1
Entangled transformer for image captioning.	MSCOCO	77.3	-	-	37.1	28.2	117.9	57.1	21.4
Meshed memory transformer for image captioning.	MSCOCO	80.8	-	-	39.1	29.2	131.2	58.6	22.6

Table 1: Results of some of the Architectures discussed. B1, B2, B3, B4 corresponds to BLEU-1,BLEU-2,BLEU-3,BLEU-4 respectively. MT stands for METEOR Score, CD stands for CIDEr Score, RG stands for ROUGE Score and SP stands for SPICE. The scores correspond to the values mentioned in the papers.

5. Future Improvements

As we can see that the evaluation results of some of the recently proposed models such as transformer models have outperformed the previous model architectures such as the Encoder - Decoder models, Making use of transformer architectures will probably be the primary focus of future image-captioning as well as Visual Q-A systems. This does not mean that the Encoder - Decoder models cannot be used. We can still try improving them by using more complex CNNs such as the DenseNets, NASNets and EfficientNet. These CNNs can be used to extract more number of features compared to VGGNets or Inception architectures. Similarly by increasing the number of layers in the LSTM or RNN Decoders can improve the performance and obtain more information. Addition of more control gates to the attention mechanism models might increase performance. Regarding other approaches,

One approach can be using Multi-Modal networks which can be used to extract multiple attributes such as geometric positions, object relationships, background information, etc. and input these to a model which can predict more accurate captions based on information received. Another approach to image captioning is that by using object detectors or recognition systems to predict and interpret all the objects present in an image and create a list of object tags. We can then pass these tags to a sentence generator model built using Natural Language Processing Toolkits or Deep Learning Models. As an additional input to these tags a relationship tag can be passed to the NLP Processor to generate more accurate sentences.

6. Conclusion

In this paper we discussed about the various Image captioning models which make use of Deep Learning architectures. Since the Encoder - Decoder architecture is the most common one we have discussed about few of them. In recent studies and research on sequence related tasks, It has been shown that Transformer architectures have produced better results than RNNs or LSTM architectures. Thus, We have mostly discussed about few Transformer based Image Captioning architectures.

Then, we have discussed about the different datasets used for Image Captioning followed by the different Evaluation Metrics used. Finally we have discussed about what all could be the possible methods, model and approaches that can be used to improve image captioning tasks.

References

[1] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822.

[2] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[3] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.

[4] Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2019). Meshed-memory transformer for image captioning.

[5] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[6] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[7] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

[8] Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

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

[10] Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR.

[11] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Li, F. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

[12] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25.

[13] Li, G., Zhu, L., Liu, P., and Yang, Y. (2019a). Entangled transformer for image captioning. In *The IEEE International Conference on Computer Vision (ICCV)*.

[14] Li, J., Yao, P., Guo, L., and Zhang, W.-C. (2019b). Boosted transformer for image captioning. *Applied Sciences*, 9:3260.

[15] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[16] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[17] Liu, C., Mao, J., Sha, F., and Yuille, A. L. (2016a). Attention correctness in neural image captioning. *CoRR*, abs/1605.09553.

[18] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. (2016b). Optimization of image description metrics using policy gradient methods. *CoRR*, abs/1612.00370.

[19] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[20] Lu, J., Xiong, C., Parikh, D., and Socher, R. (2016). Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*, abs/1612.01887.

[21] Mun, J., Cho, M., and Han, B. (2016). Text-guided attention model for image captioning. *CoRR*, abs/1612.03557.

[22] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. pages 311–318.

[23] Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

[24] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.

[25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

[26] Tran, A., Mathews, A., and Xie, L. (2020). Transform and tell: Entity-aware news image captioning. *CVPR*.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[28] Vedantam, R., Zitnick, C. L., and Parikh, D. (2014). Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

[29] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

[30] Wang, Q. and Chan, A. B. (2018). CNN+CNN: convolutional decoders for image captioning. *CoRR*, abs/1805.09019.

[31] Xu, H. and Saenko, K. (2015). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234.

[32] You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. *CoRR*, abs/1603.03925.

[33] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.