applied sciences

MDPI

Review

A Systematic Literature Review on Image Captioning

Raimonda Staniūtė and Dmitrij Šešok *

Department of Information Technologies, Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania; raimonda.staniute@stud.vgtu.lt

* Correspondence: dmitrij.sesok@vgtu.lt

Received: 26 March 2019; Accepted: 9 May 2019; Published: 16 May 2019

Abstract: Natural language problems have already been investigated for around five years. Recent progress in artificial intelligence (AI) has greatly improved the performance of models. However, the results are still not sufficiently satisfying. Machines cannot imitate human brains and the way they communicate, so it remains an ongoing task. Due to the increasing amount of information on this topic, it is very difficult to keep on track with the newest researches and results achieved in the image captioning field. In this study a comprehensive Systematic Literature Review (SLR) provides a brief overview of improvements in image captioning over the last four years. The main focus of the paper is to explain the most common techniques and the biggest challenges in image captioning and to summarize the results from the newest papers. Inconsistent comparison of results achieved in image captioning was noticed during this study and hence the awareness of incomplete data collection is raised in this paper. Therefore, it is very important to compare results of a newly created model produced with the newest information and not only with the state of the art methods. This SLR is a source of such information for researchers in order for them to be precisely correct on result comparison before publishing new achievements in the image caption generation field.

Keywords: image caption generation; NLP; LSTM; semantics; systematic literature review

1. Introduction

Ever since researchers started working on object recognition in images, it became clear that only providing the names of the objects recognized does not make such a good impression as a full human-like description. As long as machines do not think, talk, and behave like humans, natural language descriptions will remain a challenge to be solved. There have been many variations and combinations of different techniques since 2014—the very first application of neural networks in image captioning is in ref. [1]. Four successful articles [2–5], which now are the most cited articles researchers rely on, were published in 2015. There was not much interest in this area in 2014 and 2015, but it is clear from this review how exponentially the popularity is growing—57 articles found were published in 2017–2018 and already 17 were published during the first three months of 2019. The advantages and the future of human-like technologies are undoubtable; from enabling computers to interact with humans, to specific applications for child education, health assistants for the elderly or visually disabled people, and many more. While having so many opportunities for meaningful applications in society, not surprisingly many studies have already tried to obtain more accurate descriptions and make machines think like humans. However, machines still lack the natural way of human communication and this continues to be a challenging task to tackle. Our work is meant to summarize the newest articles and to give insight on the latest achievements and the highest number of results to ease the work of new researchers who would like to utilize their efforts to build better methods. This paper is a systematic literature review (SLR) of the newest articles in order to provide a summarized understanding of what has been achieved in this field so far and which techniques have

performed the best. Special attention was given to result collection and year to year comparison. We hope this work will help further researchers to find more innovative and newer ways to achieve better results. The following paper has been divided into four additional parts. First, we present the research methods which have been used to make this SLR. Second, we introduce readers to summarized tables of all the articles and results achieved in them. The purpose of the discussion section is to introduce readers to the most popular methodologies and innovative solutions in image captioning. Finally, the paper is concluded with some open questions for future studies.

2. SLR Methodology

The SLR has become a great help in the dynamic, data driven world of today, with massive data volume growth. It is sometimes very difficult to consume all currently existing information before starting to delve into a specific field. In this case, when we talk about image captioning and, as already said, having so much meaning in this task, it was found that there is much literature, which is hard to summarize and thus stay up to date with the newest achievements. There are only a few SLRs that have been conducted for image captioning until now [6–9], though with such fast progress and increasing popularity in this field we find it necessary that they continue to be undertaken. Moreover, results of image captioning models in previous reviews were not as detailed as they are in this paper. Researchers dedicated time to detailed study of most articles in image captioning—digital libraries, which store most of the articles, were identified, search questions carefully formulated, all articles found were precisely analyzed, and results presented together with important challenges which were captured through the review process. This work follows ref. [6] as a guideline due to the easily understandable structure of their work and the similar ideas.

2.1. Search Sources

Digital libraries today are the most suitable platforms for books, journals, and articles search. In this literature review we chose three digital libraries due to limited resources and the huge number of articles under this topic. However, we can clearly see that these libraries cover a significant amount of the relevant literature sources for our study. Three different digital libraries were used to execute a research:

1. ArXiv
2. IEEE Xplore
3. Web of Science—WOS (previously known as Web of Knowledge)

There have been many researches done in the field of image captioning so we narrowed down the literature review by searching for articles only from the last four years—from 2016 to 2019. During the research in the digital library, we filtered out articles, which were posted under the computer science topic.

2.2. Search Questions

It is very important to have clear questions which need to be answered after the whole literature has been reviewed. The results retrieved after each query must be precise, without too much noise and without unnecessary articles, so the questions were carefully formulated after many attempts. In this paper we answer four questions:

1. What techniques have been used in image caption generation?
2. What are the challenges in image captioning?
3. How does the inclusion of novel image description and addition of semantics improve the performance of image captioning?
4. What are the newest researches on image captioning?

Questions were selected to fully cover the main objectives of this paper—to present the main techniques used in image captioning in the past four years, as well as to identify the main challenges the researchers have faced. Furthermore, we aimed to summarize results from the newest papers published for a fair comparison of upcoming papers, so we included a generic query for image captioning, but filtered out articles from year 2019. In general, an image captioning query would be too broad and as we have a strong focus to introduce readers to the newest achievements, we need to read only the current newest articles. Although there have been a lot of good results achieved in earlier years, we omitted questions 1–3 covering the years 2016–2019. It is necessary to compare new researches with the best results achieved in image captioning which might be hard to find due to a large number of articles in this area and the low visibility of the less cited ones.

2.3. Search Query

To become acquainted with the "image caption generation" topic we first conducted a quick review of articles under it. We obtained an idea of the technologies and models, which are popular under this topic so that our research would be relevant and correct. Moreover, we did not narrow the search query to small details in order to get enough results and an appropriate number of articles from the search—keywords are presented identically to how they were submitted for the search query. The query questions together with the number of articles found in each library are presented below in Tables 1–4. Libraries were read in the order as they are listed in the table—first the query was searched in ArXiv, then in IEEE Xplore, and finally in WOS. If the article found had already been previewed from the previous library or from a previous query, it was not added to the total number of relevant articles but identified in brackets.

Table 1. Results from the search after "Techniques image caption generation" search query.

1Q	ArXiv	IEEE	WOS
Found	29	9	18
Relevant	11	6 (1)	10(6)

Table 2. Results from the search after "Challenges image caption generation" search query.

2Q	ArXiv	IEEE	WOS
Found	48	19	78
Relevant	16 (3)	3 (6)	12 (11)

Table 3. Results from the search after "Novel semantic image captioning" search query.

3Q	ArXiv	IEEE	WOS
Found	26	28	42
Relevant	6 (4)	8 (6)	2 (10)

Table 4. Results from the search after "Image captioning" search query with filtered date to 2019 as we aimed for the newest articles published.

4Q	ArXiv	IEEE	WOS
Found	38	20	25
Relevant	8 (3)	5 (3)	7 (3)

It is quite clear that the WOS digital library usually brings out the largest number of results, though with the smallest percentage of relevant articles for the topic of interest. ArXiv was the most precise and had the best ratio between relevant and all other articles from this study experience.

3. Results

After reading all the articles and inspired by another SLR [6] we achieved a good understanding on the key aspects in image captioning. To present the summarized results in a convenient way, a comprehensive comparison table (Table A1 in Appendix A [10–87]) of all articles found with the methods used was made together with the results on the most used datasets for testing. The structure is presented below:

- Year
    - 2016;
    - 2017;
    - 2018;
    - 2019;
- Feature extractors:
    - AlexNet;
    - VGG-16 Net;
    - ResNet;
    - GoogleNet (including all nine Inception models);
    - DenseNet;
- Language models:
    - LSTM;
    - RNN;
    - CNN;
    - cGRU;
    - TPGN;
- Methods:
    - Encoder-decoder;
    - Attention mechanism;
    - Novel objects;
    - Semantics;
- Results on datasets:
    - MS COCO;
    - Flickr30k;
- Evaluation metrics:
    - BLEU-1;
    - BLEU-2;
    - BLEU-3;
    - BLEU-4;
    - CIDEr;
    - METEOR;

Under each column, representing one aspect, x was written if this aspect appeared in the article. The following columns represent evaluation metrics results on two datasets —MS COCO and Flickr30k. If no testing was performed on one of the two selected datasets, the cells were left empty. If a different

dataset or evaluation metric was used in the article, a short note was provided. i.e., Ref. [10] used the Lifelog Dataset for model training and evaluation, Ref. [20]—the Visual Genome Dataset, Ref. [39] evaluated results using F-1 score metrics, Ref. [47]—R metrics. To present the results to be easier understood, we first presented five articles from each one which achieved the most results—Tables 5–8. There were only three articles from 2016 which were evaluated on MS COCO, so only those were presented. Other tables have five articles with the top five results on MS COCO based on the highest CIDEr metric results.

The distribution of each year's results based on the six main metrics is presented in Figures 1–6. The figures do not provide information on how the results changed throughout the year, yet we can still identify inconsistency from one year to another. For example, results achieved in 2018 are many times lower than the ones achieved in 2017 in all metrics. Moreover, from Tables 7 and 8 we can see that there were some results in 2018, which were higher than the results in 2019 which confirms the assumption of this study about the difficulty in keeping up with the newest articles. The highest result for the CIDEr evaluation metric on MS COCO from all articles found during this SLR was reached in 2019 in ref. [87], but was only 0,2 higher than the result from 2018 in ref. [70]. None of the papers which were published in 2019 included this result as a comparison with their achieved results. In most of the papers models were compared with state of the art methods and so they were stated to have achieved better results while there were already much higher results in different papers.

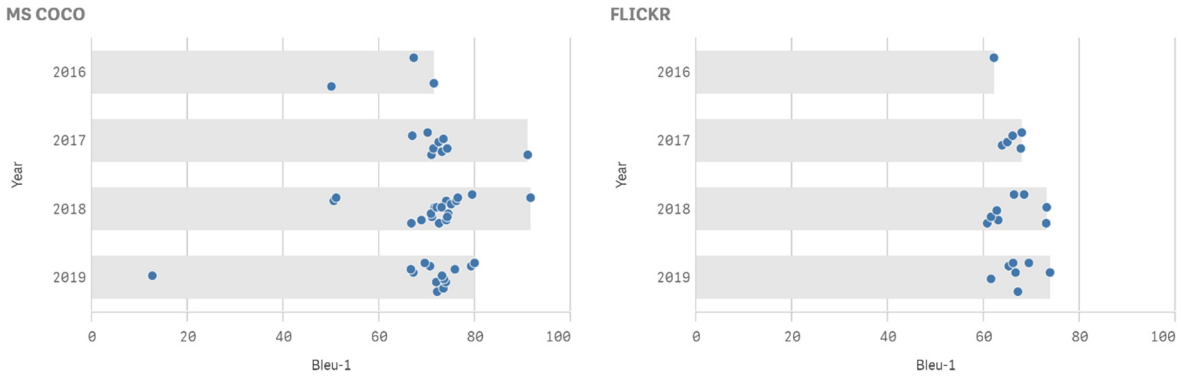

Figure 1. Bleu-1 results by year on MSCOCO (**left**) and Flickr30k (**right**).

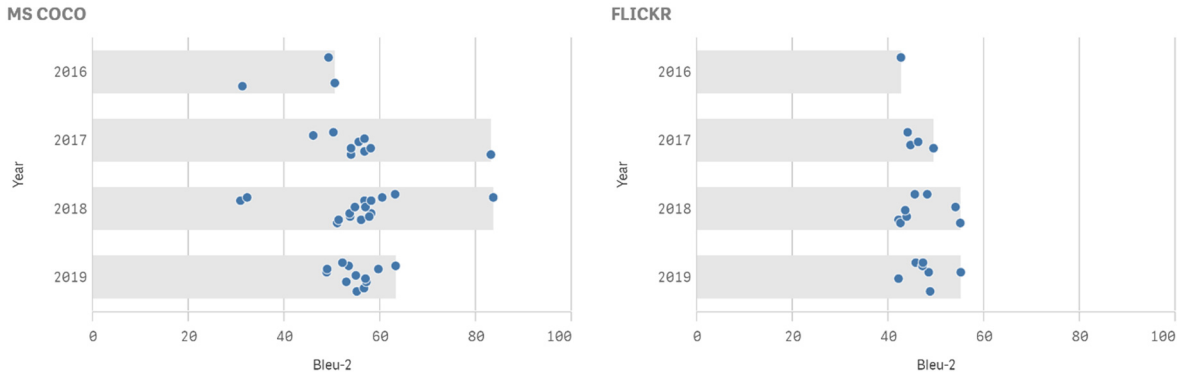

Figure 2. Bleu-2 results by year on MSCOCO (**left**) and Flickr30k (**right**).

Table 5. Top 3 results in 2016 (sorted by CIDEr result on MSCOCO).

		Image Encoder					Image Decoder					Method			MS COCO						Flickr30k					
Year	**Citation**	**AlexNet**	**VGGNet**	**GoogleNet**	**ResNet**	**DenseNet**	**LSTM**	**RNN**	**CNN**	**cGRU**	**TPGN**	**Encoder-Decoder**	**Attention**	**Novel objects**	**Semantics**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**
2016	[13]		x				x					x	x			71.4	50.5	35.2	24.5	63.8	21.9						
2016	[12]			x			x					x			x	50.0	31.2	20.3	13.1	61.8	16.8						
2016	[11]	x	x				x					x				67.2	49.2	35.2	24.4			62.1	42.6	28.1	19.3		

Table 6. Top 5 results in 2017 (sorted by CIDEr result on MSCOCO).

		Image Encoder					Image Decoder					Method			MS COCO						Flickr30k					
Year	**Citation**	**AlexNet**	**VGGNet**	**GoogleNet**	**ResNet**	**DenseNet**	**LSTM**	**RNN**	**CNN**	**cGRU**	**TPGN**	**Encoder-Decoder**	**Attention**	**Novel Objects**	**Semantics**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**
2017	[34]				x		x					x	x			74.2	58	43.9	33.2	108.5	26.6	67.7	49.4	35.4	25.1	53.1	20.4
2017	[33]				x		x					x	x			73.1	56.7	42.9	32.3	105.8	25.8						
2017	[32]		x		x		x	x				x	x			91	83.1	72.8	61.7	102.9	35						
2017	[31]		x				x					x						39.3	29.9	102	24.8			37.2	30.1	76.7	21.5
2017	[30]		x		x		x											42	31.9	101.1	25.7			32.5	22.9	44.1	19

Table 7. Top 5 results in 2018 (sorted by CIDEr result on MSCOCO).

		Image Encoder					Image Decoder					Method				MS COCO						Flickr30k					
Year	Citation	AlexNet	VGGNet	GoogleNet	ResNet	DenseNet	LSTM	RNN	CNN	cGRU	TPGN	Encoder-Decoder	Attention	Novel Objects	Semantics	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	Meteor	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	Meteor
2018	[70]				x		x					x	x						38.1	126.1	28.3						
2018	[69]				x		x					x	x						38.3	123.2	28.6						
2018	[68]				X		X					X	X	X	x	79.4	63.1	48.2	36.1	119.6	28.1	72.1	48.7	36.9	21.2	53.6	20.5
2018	[67]			x	x		x					x	x	x		76.4	60.4	47.9	37	112.5	27.4						
2018	[66]				x		x					x	x		x	76.1	58.1	44.9	34.9	109.1	26.7	73.1	54	38.6	27.9	59.4	21.7

Table 8. Top 5 results in 2019 (sorted by CIDEr result on MSCOCO).

		Image Encoder					Image Decoder					Method				MS COCO						Flickr30k					
Year	Citation	AlexNet	VGGNet	GoogleNet	ResNet	DenseNet	LSTM	RNN	CNN	cGRU	TPGN	Encoder-Decoder	Attention	Novel Objects	Semantics	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	Meteor	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	Meteor
2019	[87]				x		x					x	x	x	x			38.6	28.3	126.3							
2019	[86]				x		x					x	x	x	x	79.9			37.5	125.6	28.5	73.8	55.1	40.3	29.4	66.6	23
2019	[85]				x		x					x	x	x	x	79.2	63.2	48.3	36.3	120.2	27.6						
2019	[84]				x		x								x	75.8	59.6	46	35.6	110.5	27.3						
2019	[83]		x		x		x					x	x	x	x				55	110.1	26.1						

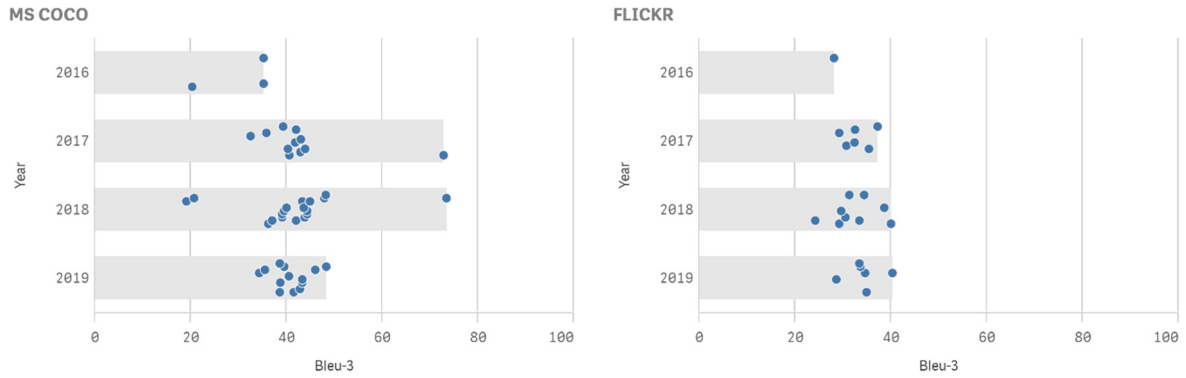

Figure 3. Bleu-3 results by year on MSCOCO (**left**) and Flickr30k (**right**).

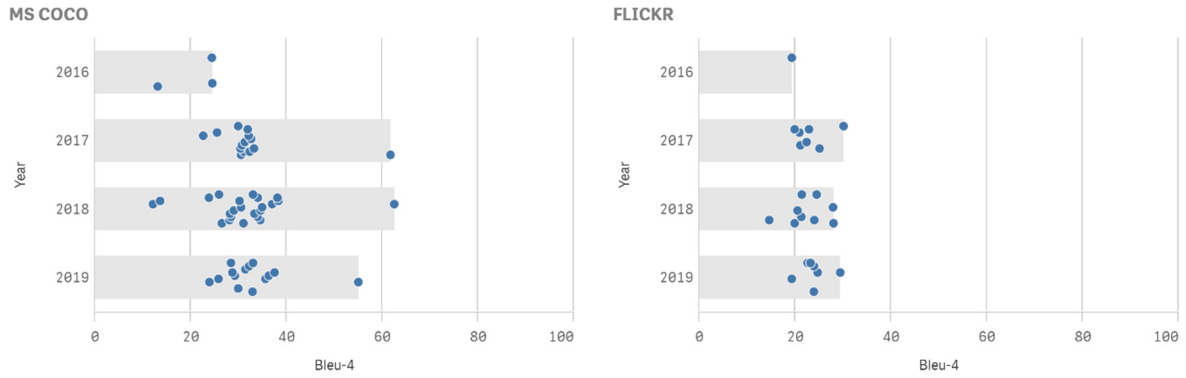

Figure 4. Bleu-4 results by year on MSCOCO (**left**) and Flickr30k (**right**).

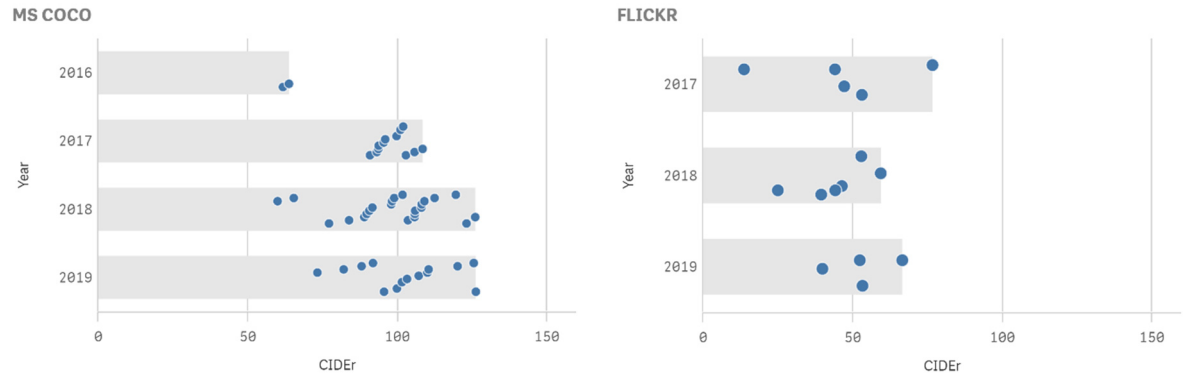

Figure 5. CIDEr results by year on MSCOCO (**left**) and Flickr30k (**right**).

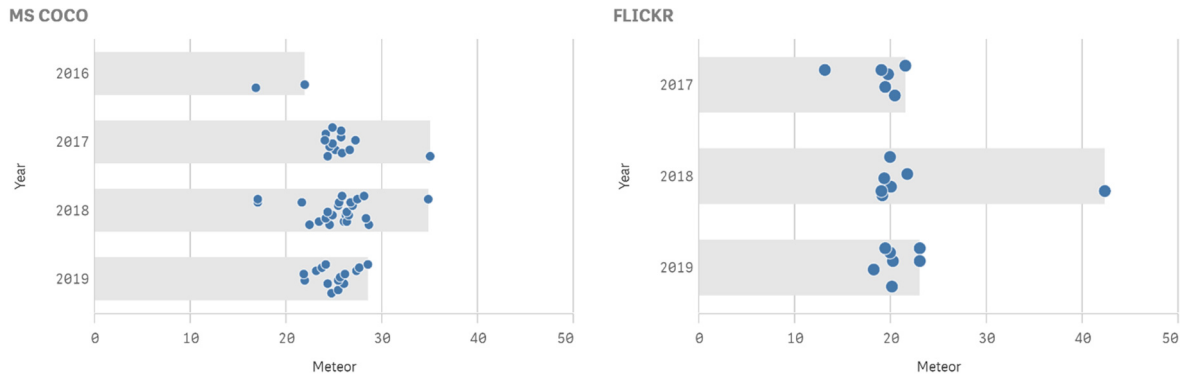

Figure 6. Meteor results by year on MSCOCO (**left**) and Flickr30k (**right**).

Tables 9 and 10 present results based on different techniques used in image captioning—which combinations of encoder and decoder were used every year and which methods were the most popular. Those tables help to understand which techniques work best together, and which combinations have probably not been successful or have not been explored at all up to now.

Table 9. Encoder/decoder results by year.

Encoder/Decoder	Year	LSTM	RNN	CNN	cGRU	TPGN
AlexNet	2016	[11]				
	2017	[22]	[45]	[21]		
	2018					
	2019					
VGG Net	2016	[10,11,13]				
	2017	[14–16,19,20,25,30–32]	[26,32]	[21]		
	2018	[40–43,47–49,54,61]		[51,52]		
	2019	[71–75,77,83]				
Google Net	2016	[12]				
	2017	[17,18,28,29]				
	2018	[43,57,60,62,67]	[50]			
	2019	[76,77,80]				
ResNet	2016					
	2017	[23,27,30,32–34]	[32]			[23]
	2018	[37,39,43,46,48,49,53,55,58,61,70]			[56,59]	
	2019	[78,83–87]	[81]			
DenseNet	2016					
	2017					
	2018					
	2019	[79]				

Table 10. Methods used by year.

Year/Method	Encoder-Decoder	Attention Mechanism	Novel Objects	Semantics
2016	[11–13]	[13]		[12]
2017	[15,17,19,22,23,25–29,31–34]	[19,21,22,26,27,32–34]	[15,18,25,27]	[27,28]
2018	[35,37,39–42,44–46,48,50–52, 54,56,57,59–64,66–70]	[36–38,40,41,46–51,53, 56,59–70]	[39,40,43–45,47–50,58, 67,68]	[36,42,44,47,50,57,58, 66,68]
2019	[71–73,75,76,78,82,83,85–87]	[71,74–81,83,85–87]	[72–75,77–80,83,85–87]	[75,80,82–87]

4. Discussion

In this paragraph we discuss the key aspects of all the papers reviewed during SLR. We also present new ideas which could possibly lead to a better image captioning performance. Each aspect is explained in a separate paragraph of this section.

4.1. Model Architecture and Computational Resources

Most of the models rely on the widespread encoder–decoder framework, which is flexible and effective. Sometimes it is defined as a structure of CNN + RNN. Usually a convolutional neural network (CNN) represents the encoder, and a recurrent neural network (RNN) the decoder. The encoder is the one which "reads" an image—given an input image, it extracts a high-level feature representation. The decoder is the one which generates words—given the image representation from the encoder (encoded image), it generates words to represent the image with a full grammatically and stylistically correct sentence.

4.1.1. Encoder—CNN

As there is usually only one encoder in the model, the performance is highly reliant on the CNN deployed. Even though we identified five convolutional networks in our research, there are two which stand out and were used the most. The first most popular choice for the feature extractor from images is VGGNet, preferred for the simplicity of the model and for its power. During this study it was found that VGG was used in 33 of 78 reviewed articles. However, the same number of articles which used ResNet as an encoder was also found. ResNet wins for being computationally the most efficient compared to all other convolutional networks. In ref. [88] a clear comparison of four networks—AlexNet, VGGNet, ResNet, and GoogleNet (also called Inception-X Net) was made—results are presented in the Table 11 below.

Table 11. Table of comparison for CNN architectures (from ref. [88]).

CONVOLUTIONAL NEURAL NETWORKS ARCHITECTURES					
Architecture	**#Params**	**#Multiply-Adds**	**Top-1 Accuracy**	**Top-5 Accuracy**	**Year**
Alexnet	61M	724M	57.1	80.2	2012
VGG	138M	15.5B	70.5	91.2	2013
Inception-V1	7M	1.43B	69.8	89.3	2013
Resnet-50	25.5M	3.9B	75.2	93	2015

It is clear from Table 11 that ResNet performs best—from both Top-1 and Top-5 accuracy. It also has much fewer parameters than VGG which saves computational resources. However, being easy to implement, VGG remains popular among researchers and has the second highest result, regarding the review from ref. [88]. The newest research mostly focuses on prioritizing simplicity and speed at a slight cost in performance. It is a matter for a researcher to decide if he or she needs more precision in the results, a more effectively performing model, or more simplicity.

4.1.2. Decoder—LSTM

LSTM (long-short-term memory) was developed from RNN, with the intention to work with sequential data. It is now considered as the most popular method for image captioning due to its effectiveness in memorizing long term dependencies through a memory cell. Undoubtedly this requires a lot of storage and is complex to build and maintain. There have been intentions to replace it with CNN [52], but as we can see from the number of times this method is used in most of the articles found during this SLR (68 of 78), scientists always come back to LSTM. LSTM works by generating a caption by making one word at every time step conditioned on a context vector, together with the previous hidden state and the earlier generated words.

Computational speed not only depends on the feature detection model, but also on the size of the vocabulary—each new word added consumes more time. Just recently [73] scientists have tried to solve the image captioning task by resizing the vocabulary dictionary. Usually the vocabulary size might vary from 10,000 to 40,000 words, while their model relies on 258 words. The decrease is quite sharp—reduced by 39 times if compared to 10,000, but the results are high, with some space for improvements.

4.1.3. Attention Mechanism

The attention model was established with an intention to replicate natural human behavior—before summarizing an image, people tend to pay attention to specific regions of that image and then form a good explanation of the relationship of objects in those regions. The same approach is used in the attention model. There are several ways in which researchers have tried to duplicate it, which are widely known as hard or soft attention mechanisms [5]. Some other scientists have highlighted top-down and bottom-up attention models. Ref. [89] recently confirmed that the better approach is still top-down attention mechanisms as the results from experiments with humans and with machines

showed similar results. In the top down model, the process starts from a given image as input and then converts it into words. Moreover, a new multi-modal dataset is created with the highest number of new instances from human fixations and scene descriptions.

4.2. Datasets

Most of the works are evaluated on Flickr30k [90] and MSCOCO [91] datasets. Both datasets are rich in the number of images and each image has five captions assigned which makes it very suitable to train and test the models. It is of course necessary to continuously compare models with the same datasets in order to check the performance, however, they are very limited in the object classes and scenarios presented. The need of new datasets has always been an open question in image captioning. Ref. [92] proposed a method for gathering large datasets of images from the internet which might be helpful for replacing MS COCO or Flickr datasets which were used in most of the previous researches. There have been several other datasets used for model evaluation, such as Lifelog dataset [10], Visual Genome dataset [20,36], IAPRTC-12 [45], OpenImages and Visual Relationship Detection datasets [36], but they were just single cases.

Recently the popularity in novel image scenarios has grown which has increased the demand of newer datasets even more. In ref. [93] the first rigorous and large-scale data set for novel object captioning, which contains more than 500 novel object classes, was introduced. Another realistic dataset was introduced in ref. [94]. It contains news images and their actual captions, along with their associated news articles, news categories, and keyword labels. Moreover, it is clear, that social networks are highly integrated into people's lifestyle. There are more and more images appearing on the social media, especially from the young generation, so it is important to analyze this data as well—for the most natural background, for the newest trends to be interpreted by machines, and to start learning and improving on those as well. Ref. [95] proposed a novel deep feature learning paradigm based on social collective intelligence, which can be acquired from the inexhaustible social multimedia content on the Web, particularly largely social images and tags, however, it was not further continued, at least to our knowledge.

4.3. Human-Like Feeling

In the last year, two keywords have come into the vocabulary of almost every article written under the image captioning topic—novel and semantics. These keywords are important for solving the biggest challenge in this exercise i.e. generating a caption in a way that it would be inseparable from human written ones. Semantics implementation [49] is supposed to design a clean way of injecting sentiment into the current image captioning system. Novel objects must be included for the expansion of scenarios. There have been several insights on why this is still an open issue. First of all, usually models are built on very specific datasets, which do not cover all possible scenarios and are not applicable in describing diverse environment. The same with vocabulary as it has a limited number of words and their combinations. Second, models are usually thought to perform on one specific task, while humans are able to work on many tasks simultaneously. Ref. [35] has already tried to overcome this problem and has provided a solution although it was not further continued. Another great approach for dealing with unseen data, as it is currently impossible to feed all existing data into the machine, was proposed in ref. [56,96]. Lifelong learning is based on a questioning approach i.e. making a discussion directly with the user or inside the model. This approach relies on a natural way of human communication; from early childhood children mostly learn by asking questions. The model is intended to learn also like a child—by asking specific questions and learning from the answers. This method falls under the question answering topic—a literature research in depth might be done on this topic as here we have presented only what appeared during this study on image captioning. This can be targeted as a separate problem, but it also makes a great impact in image captioning.

4.4. Comparison of Results

This study found many articles in which the results of their models had been compared with state of the art models, such as refs. [2–5]. As these models were built some years ago, they have been more cited so are easier to find during a search on the digital libraries. For example, ref. [5] has been cited 2855 times, according to Google Scholar from Google, while most of the newest articles found have not been cited at all yet, or the ones written in 2018 have usually been cited less than 10 times. Not surprisingly the newer the articles are, the further at the bottom of the search they appear, so most researchers might not even find them if not enough time has been dedicated for a literature review. Figures 1–6 confirm that results are not steadily increasing—there are many results which are not higher than the ones from a year ago. This can undoubtedly be due to the topic difficulty, but also lack of details can lower the goals of researchers so they do not improve knowing that there are higher results already even though a very important part for researchers is to compare their work results with similar approaches. In this study the results from the newest models are presented so upcoming researchers can compare their models with regard to the newest achievements. We hope this research will help further researchers to save their time on detailed literature reviews and to keep in mind the importance of checking for the newest articles.

5. Conclusions

Image captioning is a very exciting exercise and raises tough competition among researchers. There are more and more scientists who are deciding to explore this study field, so the amount of information is constantly increasing. It was noticed that the results are usually compared with quite old articles, although there are dozens of new ones, with even higher results and new ideas for improvements. The comparison with older articles gives a misunderstanding of the real view of result increase—usually there have been much higher results already achieved, however not included in the paper. New ideas can also very easily become lost if they are not looked for carefully. In order to prevent good ideas been lost and to increase fair competition among the new models created, this systematic literature review summarizes all the newest articles and their results in one place. Moreover, it is still not clear if MS COCO and Flick30k datasets are enough for model evaluation and if they serve sufficiently well when having in mind diverse environments. The amount of data will never stop increasing and new information will keep appearing, so future studies should consider if static models are good enough when thinking of long term application or if lifelong learning should be increasingly thought of. We hope this SLR will serve other scientists as a guideline and as an encouragement of the newest information to be collected for their research evaluation.

Author Contributions: All authors contributed to designing and performing measurements, data analysis, scientific discussions, and writing the article.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Comprehensive comparison table of all articles from the study.

		Image Encoder					Image Decoder					Method				MS COCO						Flickr30k					
Year	**Citation**	**AlexNet**	**VGGNet**	**GoogleNet**	**ResNet**	**DenseNet**	**LSTM**	**RNN**	**CNN**	**cGRU**	**TPGN**	**Encoder-Decoder**	**Attention**	**Novel Objects**	**Semantics**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**
2016	[10]		x				x									Lifelog Dataset											
2016	[11]	x	x				x					x				67.2	49.2	35.2	24.4			62.1	42.6	28.1	19.3		
2016	[12]			x			x					x			x	50.0	31.2	20.3	13.1	61.8	16.8						
2016	[13]		x				x					x	x			71.4	50.5	35.2	24.5	63.8	21.9						
2017	[14]		x				x																				
2017	[15]		x				x					x		x								63.8	44.6	30.7	21.1		
2017	[16]		x				x									CUB-Justify											
2017	[17]			x			x					x									27.2						
2017	[18]			x			x							x		66.9	46	32.5	22.6								
2017	[19]		x				x					x	x			70.1	50.2	35.8	25.5		24.1	67.9	44	29.2	20.9		19.7
2017	[20]		x				x									Visual Genome Dataset											
2017	[21]	x	x										x												19.9	13.7	13.1
2017	[22]	x					x					x	x														
2017	[23]				x		x				x	x				70.9	53.9	40.6	30.5	90.9	24.3						
2017	[24]						x												31.1	93.2							
2017	[25]		x				x					x		x		71.3	53.9	40.3	30.4	93.7	25.1						
2017	[26]		x					x				x	x						30.7	93.8	24.5						
2017	[27]				x		x					x	x	x	x	72.4	55.5	41.8	31.3	95.5	24.8	64.9	46.2	32.4	22.4	47.2	19.4
2017	[28]			x			x					x			x	73.4	56.7	43	32.6	96	24						
2017	[29]			x			x					x							32.1	99.8	25.7	66					
2017	[30]		x		x		x											42	31.9	101.1	25.7			32.5	22.9	44.1	19
2017	[31]		x				x					x						39.3	29.9	102	24.8			37.2	30.1	76.7	21.5
2017	[32]		x		x		x	x				x	x			91	83.1	72.8	61.7	102.9	35						
2017	[33]				x		x					x	x			73.1	56.7	42.9	32.3	105.8	25.8						
2017	[34]				x		x					x	x			74.2	58	43.9	33.2	108.5	26.6	67.7	49.4	35.4	25.1	53.1	20.4
2018	[35]						x					x				Recall Evaluation metric											
2018	[36]												x		x	OI, VG, VRD											
2018	[37]				x		x					X	X			71.6	51.8	37.1	26.5		24.3						
2018	[38]						x						x			APRC, CSMC											
2018	[39]				x		x					x		x		F-1 score metrics					21.6						

Table A1. *Cont.*

		Image Encoder					Image Decoder					Method				MS COCO						Flickr30k					
Year	**Citation**	**AlexNet**	**VGGNet**	**GoogleNet**	**ResNet**	**DenseNet**	**LSTM**	**RNN**	**CNN**	**cGRU**	**TPGN**	**Encoder-Decoder**	**Attention**	**Novel Objects**	**Semantics**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**
2018	[40]		x				x					x	x	x													
2018	[41]		x				x					x	x			72.5	51	36.2	25.9		24.5	68.4	45.5	31.3	21.4		19.9
2018	[42]		x				x					x			x	74	56	42	31		26	73	55	40	28		
2018	[43]		x	x	x		x							x								63	42.1	24.2	14.6	25	42.3
2018	[44]						x					x		x	x	74.4	58.1	44.3	33.8		26.2						
2018	[45]	x						x				x		x				IAPRTC-12									
2018	[46]				x		x					x	x														
2018	[47]		x				x						x	x	x	R											
2018	[48]		x		x		x					x	x	x		50.5	30.8	19.1	12.1	60	17						
2018	[49]		x		x		x						x	x		51	32.2	20.7	13.6	65.4	17						
2018	[50]			x				x				x	x	x	x	66.7			23.8	77.2	22.4						
2018	[51]		x						x			x	x			68.8	51.3	37	26.5	83.9	23.4	60.7	42.5	29.2	19.9	39.5	19.1
2018	[52]		x						x			x				71	53.7	39.1	28.1	89	24.1						
2018	[53]				x		x						x			70.8	53.6	39.1	28.4	89.8	24.8	61.5	43.8	30.5	21.3	46.4	20
2018	[54]		x				x					x						39.5	28.2	90.7	24.3						
2018	[55]				x		x									72	54.7	40	29	91.7		62.7	43.5	29.6	20.5		19.3
2018	[56]				x					x		x	x				57.1		30.5	98	25.4						
2018	[57]			x			x					X			x	74.0	56.7	43.3	31.3	98.3	25.5	64.6	43.8	31.9	22.4	39.6	19.2
2018	[58]				x		x							x	x	91.6	83.6	73.4	62.5	99	34.8						
2018	[59]				x					x		x	x						30.2	101.8	25.8						
2018	[60]			x			x					x	x						34	103.6	26.3						
2018	[61]		x		x		x					x	x			74.2	57.7	43.8	33	105.8		66.3	48.1	34.4	24.5	52.9	
2018	[62]			x	x		x					x	x					44.3	34.5	105.9	26.5			33.4	24	44.2	19
2018	[63]				x		x					x	x					44.3	34	106.0	26.3						
2018	[64]				x		x					x	x			73	56.9	43.6	33.3	108.1							
2018	[65]				x		x						x			75			34.6	108.2	26.9						
2018	[66]				x		x					x	x		x	76.1	58.1	44.9	34.9	109.1	26.7	73.1	54	38.6	27.9	59.4	21.7
2018	[67]			x	x		x					x	x	x		76.4	60.4	47.9	37	112.5	27.4						
2018	[68]				X		X					X	X	X	x	79.4	63.1	48.2	36.1	119.6	28.1	72.1	48.7	36.9	21.2	53.6	20.5
2018	[69]				x		x					x	x						38.3	123.2	28.6						
2018	[70]				x		x					x	x						38.1	126.1	28.3						
2019	[71]		x				x					x	x			71.9	52.9	38.7	28.4		24.3	69.4	45.7	33.2	22.6		23
2019	[72]		x				x					x		x							21.9						
2019	[73]		x				x					x		x		12.6											

Table A1. *Cont.*

		Image Encoder					Image Decoder					Method				MS COCO						Flickr30k					
Year	**Citation**	**AlexNet**	**VGGNet**	**GoogleNet**	**ResNet**	**DenseNet**	**LSTM**	**RNN**	**CNN**	**cGRU**	**TPGN**	**Encoder-Decoder**	**Attention**	**Novel Objects**	**Semantics**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**Meteor**
2019	[74]		x				x						x	x		67.1	48.8	34.3	23.9	73.3	21.8						
2019	[75]		x				x					x	x	x	x	66.6	48.9	35.5	25.8	82.1	23.1	61.5	42.1	28.6	19.3	39.9	18.2
2019	[76]			x			x					x	x			70.6	53.4	39.5	29.2	88.1	23.7						
2019	[77]		x	x			x						x	x		69.5	52.1	38.6	28.7	91.9	24.1	66.6	48.4	34.6	24.7	52.4	20.2
2019	[78]				x		x					x	x	x		72.1	55.1	41.5	31.4	95.6	24.7						
2019	[79]					x	x						x	x		73.4	56.6	42.8	32.2	99.9	25.4	65.2	47.1	33.6	23.9		19.9
2019	[80]			x			x						x	x	x	73.9	57.1	43.3	33	101.6	26	66.1	47.2	33.4	23.2		19.4
2019	[81]				x			x					x			73.5	56.9	43.3	32.9	103.3	25.4	67.1	48.7	34.9	23.9	53.3	20.1
2019	[82]						x					x			x	73.1	54.9	40.5	29.9	107.2	25.6						
2019	[83]		x		x		x					x	x	x	x				55	110.1	26.1						
2019	[84]				x		x								x	75.8	59.6	46	35.6	110.5	27.3						
2019	[85]				x		x					x	x	x	x	79.2	63.2	48.3	36.3	120.2	27.6						
2019	[86]				x		x					x	x	x	x	79.9			37.5	125.6	28.5	73.8	55.1	40.3	29.4	66.6	23
2019	[87]				x		x					x	x	x	x			38.6	28.3	126.3							

References

1. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.
2. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
3. Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning based on a hierarchical attention mechanism and policy gradient optimization. *J. Latex Cl. Files* **2015**, *14*, 8.
4. Karpathy, A.; Li, F.-F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; Volume 39, pp. 3128–3137. [CrossRef]
5. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML'15), Lille, France, 6–11 July 2015.
6. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CSUR)* **2019**, *51*, 118. [CrossRef]
7. Bai, S.; An, S. A survey on automatic image caption generation. *Neurocomputing* **2018**, *311*, 291–304. [CrossRef]
8. Shabir, S.; Arafat, S.Y. An image conveys a message: A brief survey on image description generation. In Proceedings of the 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), Azad Kashmir, Pakistan, 9–10 April 2018; pp. 1–6. [CrossRef]
9. Jenisha, T.; Purushotham, S. A survey of neural network algorithms used for image annotation. *IIOAB J.* **2016**, *7*, 236–252.
10. Fan, C.; Crandall, D.J. DeepDiary: Automatic caption generation for lifelogging image streams. In Proceedings of the European Conference on Computer Vision, Amsterdam, the Netherlands, 11–14 October 2016; pp. 459–473. [CrossRef]
11. Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, the Netherlands, 15–19 October 2016; pp. 988–997. [CrossRef]
12. Mathews, A.P.; Xie, L.; He, X. Senticap: Generating image descriptions with sentiments. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
13. Sugano, Y.; Bulling, A. Seeing with humans: Gaze-assisted neural image captioning. *arXiv* **2016**, arXiv:1608.05203.
14. Venugopalan, S.; Anne Hendricks, L.; Rohrbach, M.; Mooney, R.; Darrell, T.; Saenko, K. Captioning images with diverse objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5753–5761.
15. He, X.; Shi, B.; Bai, X.; Xia, G.-S.; Zhang, Z.; Dong, W. Image caption generation with part of speech guidance. *Pattern Recognit. Lett.* **2017**, *119*, 229–237. [CrossRef]
16. Vedantam, R.; Bengio, S.; Murphy, K.; Parikh, D.; Chechik, G. Context-aware captions from context-agnostic supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 251–260. [CrossRef]
17. Shetty, R.; Rohrbach, M.; Anne Hendricks, L.; Fritz, M.; Schiele, B. Speaking the same language: Matching machine to human captions by adversarial training. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4135–4144.
18. Xin, M.; Zhang, H.; Yuan, D.; Sun, M. Learning discriminative action and context representations for action recognition in still images. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 757–762. [CrossRef]
19. Yuan, A.; Li, X.; Lu, X. FFGS: Feature fusion with gating structure for image caption generation. In Proceedings of the Computer Vision, Tianjin, China, 11–14 October 2017; Springer: Singapore, 2017; pp. 638–649. [CrossRef]
20. Yang, L.; Tang, K.; Yang, J.; Li, L.J. Dense captioning with joint inference and visual context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2193–2202. [CrossRef]

21. Kilickaya, M.; Akkus, B.K.; Cakici, R.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. Data-driven image captioning via salient region discovery. *IET Comput. Vis.* **2017**, *11*, 398–406. [CrossRef]
22. Shah, P.; Bakrola, V.; Pati, S. Image captioning using deep neural architectures. In Proceedings of the 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 17–18 March 2017; pp. 1–4. [CrossRef]
23. Huang, Q.; Smolensky, P.; He, X.; Deng, L.; Wu, D. Tensor product generation networks for deep NLP modeling. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, LA, USA, 1–6 June 2018; pp. 1263–1273. [CrossRef]
24. Liu, C.; Sun, F.; Wang, C. MAT: A multimodal translator for image captioning. In Proceedings of the Artificial Neural Networks and Machine Learning, Pt II, Alghero, Italy, 11–14 September 2017; Lintas, A., Rovetta, S., Verschure, P., Villa, A.E.P., Eds.; Springer International Publishing Ag: Cham, Switzerland, 2017; Volume 10614, p. 784.
25. Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; Li, L.-J. Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 290–298.
26. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1242–1250.
27. Fu, K.; Jin, J.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2321–2334. [CrossRef]
28. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting image captioning with attributes. In Proceedings of the IEEE Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4894–4902. [CrossRef]
29. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 652–663. [CrossRef]
30. Xu, K.; Wang, H.; Tang, P. Image captioning with deep LSTM based on sequential residual. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 361–366. [CrossRef]
31. Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional GAN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2970–2979. [CrossRef]
32. Dai, B.; Lin, D. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*; MIT Press: London, UK, 2017; pp. 898–907.
33. Liu, C.; Sun, F.; Wang, C.; Wang, F.; Yuille, A. MAT: A multimodal attentive translator for image captioning. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 4033–4039. [CrossRef]
34. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
35. Nguyen, D.-K.; Okatani, T. Multi-task learning of hierarchical vision-language representation. *arXiv* **2018**, arXiv:1812.00500.
36. Zhang, J.; Shih, K.; Tao, A.; Catanzaro, B.; Elgammal, A. An interpretable model for scene graph generation. *arXiv* **2018**, arXiv:1811.09543.
37. Yan, S.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning using adversarial networks and reinforcement learning. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 248–253. [CrossRef]
38. Liu, D.; Fu, J.; Qu, Q.; Lv, J. BFGAN: Backward and forward generative adversarial networks for lexically constrained sentence generation. *arXiv* **2018**, arXiv:1806.08097.
39. Wu, Y.; Zhu, L.; Jiang, L.; Yang, Y. Decoupled novel object captioner. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; ACM: New York, NY, USA, 2018; pp. 1029–1037. [CrossRef]

40. Wang, W.; Ding, Y.; Tian, C. A novel semantic attribute-based feature for image caption generation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3081–3085. [CrossRef]
41. Chang, Y.-S. Fine-grained attention for image caption generation. *Multimed. Tools Appl.* **2018**, *77*, 2959–2971. [CrossRef]
42. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; van den Hengel, A. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1367–1381. [CrossRef]
43. Baig, M.M.A.; Shah, M.I.; Wajahat, M.A.; Zafar, N.; Arif, O. Image caption generator with novel object injection. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–8. [CrossRef]
44. Chen, F.; Ji, R.; Sun, X.; Wu, Y.; Su, J. GroupCap: Group-based image captioning with structured relevance and diversity constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1345–1353.
45. Kinghorn, P.; Zhang, L.; Shao, L. A region-based image caption generator with refined descriptions. *Neurocomputing* **2018**, *272*, 416–424. [CrossRef]
46. Cornia, M.; Baraldi, L.; Cucchiara, R. Show, control and tell: A framework for generating controllable and grounded captions. *arXiv* **2018**, arXiv:1811.10652.
47. Huang, F.; Zhang, X.; Li, Z.; Zhao, Z. Bi-directional spatial-semantic attention networks for image-text matching. *IEEE Trans. Image Process.* **2018**, *28*, 2008–2020. [CrossRef]
48. Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; Luo, J. "Factual" or "emotional": Stylized image captioning with adaptive learning and attention. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 527–543. [CrossRef]
49. You, Q.; Jin, H.; Luo, J. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv* **2018**, arXiv:1801.10121.
50. Mathews, A.; Xie, L.; He, X. SemStyle: Learning to generate stylised image captions using unaligned text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8591–8600. [CrossRef]
51. Wang, Q.; Chan, A.B. CNN+CNN: Convolutional decoders for image captioning. *arXiv* **2018**, arXiv:1805.09019.
52. Aneja, J.; Deshpande, A.; Schwing, A. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5561–5570.
53. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 48. [CrossRef]
54. Shi, H.; Li, P. Image captioning based on deep reinforcement learning. In Proceedings of the 10th International Conference on Internet Multimedia Computing and Service, Nanjing, China, 17–19 August 2018; p. 45.
55. Fan, Y.; Xu, J.; Sun, Y.; He, B. Long-term recurrent merge network model for image captioning. In Proceedings of the 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, Greece, 5–7 November 2018; pp. 254–259. [CrossRef]
56. Shen, K.; Kar, A.; Fidler, S. Lifelong learning for image captioning by asking natural language questions. *arXiv* **2018**, arXiv:1812.00235.
57. Yu, N.; Song, B.; Yang, J.; Zhang, J. Topic-oriented image captioning based on order-embedding. *IEEE Trans. Image Process.* **2019**, *28*, 2743–3754. [CrossRef]
58. Kim, B.; Lee, Y.H.; Jung, H.; Cho, C. Distinctive-attribute extraction for image captioning. *Eur. Conf. Comput. Vis.* **2018**, 133–144. [CrossRef]
59. Delbrouck, J.-B.; Dupont, S. Bringing back simplicity and lightliness into neural image captioning. *arXiv* **2018**, arXiv:1810.06245.
60. Sow, D.; Qin, Z.; Niasse, M.; Wan, T. A sequential guiding network with attention for image captioning. *arXiv* **2018**, arXiv:1811.00228.
61. Zhu, X.; Li, L.; Liu, J.; Li, Z.; Peng, H.; Niu, X. Image captioning with triple-attention and stack parallel LSTM. *Neurocomputing* **2018**, *319*, 55–65. [CrossRef]
62. Fang, F.; Wang, H.; Chen, Y.; Tang, P. Looking deeper and transferring attention for image captioning. *Multimed. Tools Appl.* **2018**, *77*, 31159–31175. [CrossRef]

63. Fang, F.; Wang, H.; Tang, P. Image captioning with word level attention. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018. [CrossRef]
64. Zhu, X.; Li, L.; Liu, J.; Peng, H.; Niu, X. Captioning transformer with stacked attention modules. *Appl. Sci.* **2018**, *8*, 739. [CrossRef]
65. Wang, F.; Gong, X.; Huang, L. Time-dependent pre-attention model for image captioning. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3297–3302. [CrossRef]
66. Ren, L.; Hua, K. Improved image description via embedded object structure graph and semantic feature matching. In Proceedings of the IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 10–12 December 2018; pp. 73–80. [CrossRef]
67. Jiang, W.; Ma, L.; Jiang, Y.G.; Liu, W. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 499–515.
68. Yang, M.; Zhao, W.; Xu, W.; Feng, Y.; Zhao, Z.; Chen, X.; Lei, K. Multitask learning for cross-domain image captioning. *IEEE Trans. Multimed.* **2018**, *21*, 1047–1061. [CrossRef]
69. Chen, C.; Mu, S.; Xiao, W.; Ye, Z.; Wu, L.; Ju, Q. Improving image captioning with conditional generative adversarial nets. *arXiv* **2018**, arXiv:1805.07112.
70. Du, J.; Qin, Y.; Lu, H.; Zhang, Y. Attend more times for image captioning. *arXiv* **2018**, arXiv:1812.03283.
71. Yuan, A.; Li, X.; Lu, X. 3G structure for image caption generation. *Neurocomputing* **2019**, *330*, 17–28. [CrossRef]
72. Kim, D.-J.; Choi, J.; Oh, T.-H.; Kweon, I.S. Dense relational captioning: Triple-stream networks for relationship-based captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019.
73. Mishra, A.; Liwicki, M. Using deep object features for image descriptions. *arXiv* **2019**, arXiv:1902.09969.
74. Xu, N.; Liu, A.-A.; Liu, J.; Nie, W.; Su, Y. Scene graph captioner: Image captioning based on structural visual representation. *J. Vis. Commun. Image Represent.* **2019**, *58*, 477–485. [CrossRef]
75. Tan, Y.H.; Chan, C.S. Phrase-based image caption generator with hierarchical LSTM network. *Neurocomputing* **2019**, *333*, 86–100. [CrossRef]
76. Tan, J.H.; Chan, C.S.; Chuah, J.H. COMIC: Towards a compact image captioning model with attention. *IEEE Trans. Multimed.* **2019**. [CrossRef]
77. He, C.; Hu, H. Image captioning with text-based visual attention. *Neural Process. Lett.* **2019**, *49*, 177–185. [CrossRef]
78. Li, J.; Ebrahimpour, M.K.; Moghtaderi, A.; Yu, Y.-Y. Image captioning with weakly-supervised attention penalty. *arXiv* **2019**, arXiv:1903.02507.
79. He, X.; Yang, Y.; Shi, B.; Bai, X. VD-SAN: Visual-densely semantic attention network for image caption generation. *Neurocomputing* **2019**, *328*, 48–55. [CrossRef]
80. Zhao, D.; Chang, Z.; Guo, S. A multimodal fusion approach for image captioning. *Neurocomputing* **2019**, *329*, 476–485. [CrossRef]
81. Wang, W.; Hu, H. Image captioning using region-based attention joint with time-varying attention. *Neural Process. Lett.* **2019**, 1–13. [CrossRef]
82. Zhou, Y.; Sun, Y.; Honavar, V. Improving image captioning by leveraging knowledge graphs. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 283–293. [CrossRef]
83. Ren, L.; Qi, G.; Hua, K. Improving diversity of image captioning through variational autoencoders and adversarial learning. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 263–272. [CrossRef]
84. Zhang, M.; Yang, Y.; Zhang, H.; Ji, Y.; Shen, H.T.; Chua, T.-S. More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Trans. Image Process.* **2019**, *28*, 32–44. [CrossRef]
85. Li, X.; Jiang, S. Know more say less: Image captioning based on scene graphs. *IEEE Trans. Multimed.* **2019**. [CrossRef]
86. Gao, L.; Li, X.; Song, J.; Shen, H.T. Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]
87. Zha, Z.J.; Liu, D.; Zhang, H.; Zhang, Y.; Wu, F. Context-aware visual policy network for fine-grained image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]

88. KOUSTUBH. ResNet, AlexNet, VGGNet, Inception: Understanding Various Architectures of Convolutional Networks. Available online: https://cv-tricks.com/cnn/understand-resnet-alexnet-vgg-inception/ (accessed on 24 May 2019).
89. He, S.; Tavakoli, H.R.; Borji, A.; Pugeault, N. A synchronized multi-modal attention-caption dataset and analysis. *arXiv* **2019**, arXiv:1903.02499.
90. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2641–2649. [CrossRef]
91. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
92. Ordonez, V.; Han, X.; Kuznetsova, P.; Kulkarni, G.; Mitchell, M.; Yamaguchi, K.; Stratos, K.; Goyal, A.; Dodge, J.; Mensch, A.; et al. Large scale retrieval and generation of image descriptions. *Int. J. Comput. Vis.* **2016**, *119*, 46–59. [CrossRef]
93. Agrawal, H.; Desai, K.; Chen, X.; Jain, R.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. *arXiv* **2018**, arXiv:1812.08658.
94. Tariq, A.; Foroosh, H. A context-driven extractive framework for generating realistic image descriptions. *IEEE Trans. Image Process.* **2017**, *26*, 619–632. [CrossRef]
95. Zhang, H.; Shang, X.; Luan, H.; Wang, M.; Chua, T.-S. Learning from collective intelligence: Feature learning using social images and tags. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *13*, 1. [CrossRef]
96. Sharma, S.; Suhubdy, D.; Michalski, V.; Kahou, S.E.; Bengio, Y. ChatPainter: Improving text to image generation using dialogue. *arXiv* **2018**, arXiv:1802.08216. [CrossRef]