The Lexical Gap: An Improved Measure of Automated Image Description Quality

Austin Kershaw
University of Surrey
ak00789@surrey.ac.uk

Miroslaw Bober
University of Surrey
mbober@surrey.ac.uk

April 1, 2019

Abstract

The challenge of automatically describing images and videos has stimulated much research in Computer Vision and Natural Language Processing. In order to test the semantic abilities of new algorithms, we need reliable and objective ways of measuring progress. Using our dataset of 2K human and machine descriptions, we find that standard evaluation measures alone do not adequately measure the semantic richness of a description. We introduce and test a new measure of semantic ability based on relative lexical diversity. We show how our measure can work alongside existing measures to achieve state of the art correlation with human judgement of quality.

1 Introduction

Image and video processing systems are being developed for a wide variety of semantically rich tasks, such storytelling (Zhu et al., 2015), Visual Question Answering (VQA) (Anderson et al., 2017; Teney et al., 2016; Wu et al., 2016), and engaging in visual dialogue (Jain et al., 2018). In this paper, we consider the task of Image Description (Lin et al., 2014; Hodosh et al., 2015; Plummer et al., 2017). Closing the semantic gap between human and machine descriptions requires robust and standardised measures of performance. In classical computer vision problems such as object detection, segmentation and classification, quality can be defined easily as a comparison between machine predictions and reference answers. Standard measures of image description quality consider the alignment of candidate sentences with ground truth sentences. However defining a set of "correct" answers for a given image is restrictive, as an image may contain diverse semantic information. Consequently we find semantically rich and detailed content is regarded very poorly by such measures, and the more sparse and simplistic the reference data and predictions, the higher the score. In summary:

1. We sourced 2K human and machine descriptions, which we used to show that standard automated measures of quality give an incomplete picture of semantic ability. The measures produce higher scores when candidates and reference data are semantically sparse, and lower scores on richer descriptions.

2. We show that measuring the relative lexical diversity of a system is a better indicator of semantic ability. We define two measures of relative diversity, and show that when combined with standard measures, achieve achieve state-of-the-art correlation with human judgement.

We hope our work will stimulate research in to more advanced measures of semantic ability, helping to close the gap between human and machine descriptions.

2 Relevant Literature

The predominant approach to generating original descriptions is to encode visual data into semantically useful features, which are then decoded into language. The capability of Convolutional Neural

Networks (CNN) and their variants for extracting spatial features is well established in Computer Vision. Pre-training the network on a dataset such as ImageNet[1] (which already embeds images based on the WordNet nouns contained within them) provides spatial features which accurately predict common nouns. In language generation, it is common to use a gated recurrent neural network which predicts a probability distribution across the vocabulary, given prior states and spatial features (Long et al., 2014). Many systems have evolved from this fundamental approach, and we refer interested readers to surveys on such developments(Bernardi et al., 2016; Aafaq et al., 2018). Systems are typically trained end-to-end on one of a number of image description datasets. Relevant to this paper are MS-COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2015), and Flickr30k(Plummer et al., 2017).

2.1 Methods of Evaluation

Objective measures of performance enable the automatic evaluation of systems across large datasets, avoiding the laborious process of sourcing human judgements. The measures divide into three groups:

1. Machine Translation measures: Early description systems considered image description as a translation task, in which information in the visual domain, is translated to the linguistic domain. As such machine translation measures based on n-gram alignment such as BLEU(Papineni et al., 2002), ROUGE(Lin and Hovy, 2003) and METEOR(Denkowski and Lavie, 2014).

2. Captioning Measures: CIDEr(Vedantam et al., 2015) and SPICE(Anderson et al., 2016), designed specifically for the description task. CIDEr addresses the problem of description diversity by rewarding candidates that match the consensus of references. SPICE, applies work from scene graph generation(Schuster et al., 2015) to create semantic graph representations of candidate and ground truth.

3. Neural Network Evaluation: Neural networks can be trained to evaluate descriptions. NNEVAL (Sharif et al., 2018) is a network trained to predict whether a description is human or machine, using both the captioning and translation measures as linguistic features.

As automated measures are a substitute for human evaluation, they are compared on the basis of their ability to correlate with human judgement. The poor correlation of translation measures is well known, (Bernardi et al., 2016; Chen and Dolan, 2011), and captioning measures show improved results. In this work we assess the correlation using the Composite dataset(Aditya et al., 2015). Human and machine captions for images in subsets of MS-COCO, Flickr8K and Flickr30K are judged by Amazon Mechanical Turk workers, and rated for correctness and completeness.

2.2 Lexical Diversity (LD)

The ability of text or speech to convey information specifically and articulately is a widely studied field. It is of interest in areas such as language learning, educational psychology and the study of speech impediments (Durán et al., 2004; Jarvis, 2013). An indicator of such fluency is Lexical Diversity (LD), which is a measure of the distribution of words used in a sample text. A simple measure such as the Type Token Ratio (TTR) considers the number of unique words used, relative to the total number of words in a sample. However TTR disadvantages longer texts, because for every additional word added to a corpus, the probability that it will be novel decreases. Such a measure would therefore be difficult to apply to a large scale image description corpus. A variety of measures derived from TTR have been proposed to address the issue of sample size such as the rate at which the TTR falls as successive tokens are added to the text (Jarvis, 2013). A curve with a larger negative gradient demonstrates more diversity than one with a smaller decay, and its parameters can be found with a numerical method (Durán et al., 2004). We later illustrate the application of this to image descriptions. More recent measures such as MTLD (McCarthy and Jarvis, 2010) consider the mean length of word strings for a particular TTR.

[1] http://www.image-net.org/

Hypo-geometric Distribution-D (HD-D) (McCarthy and Jarvis, 2010) measures the probability that for a random sample of words from a corpus, a particular token will be selected a certain number of times. Here we use HD-D for its simple implementation, lower sensitivity to corpus size and wide use in the literature, but our method could be applied with a different LD measure.

3 Evaluation Measures and Rich Descriptions

A desirable quality of a description is to convey semantically insightful information. In this section we describe how we sourced a set of human and machine descriptions, comparing them on their semantic richness. We compared standard evaluation measures on semantically sparse and rich captions.

3.1 Sourcing Rich and Sparse Descriptions

We showed a total of 20 images to volunteers (Figure 1), asking them to describe the image in an informative sentence. "*Describe this image as if describing it to a friend*". Unlike large scale data collection, where participants have many images to process, our smaller scale collection gave participants unlimited time to consider their description. We also sourced machine descriptions by training a common image captioning baseline(Xu et al., 2016) on MS-COCO. After validating the performance of our system against the original paper, we sourced 1K machine descriptions of our images. From a subjective comparison between the human and machine descriptions, we noted a gap in semantic richness, illustrated in Figure 2. Humans incorporate information extrinsic to the images, such as from current affairs, cultural background and human experience, reacting with empathy to emotional cues. Machine descriptions however, are produced sequentially one word at a time, with each word selected from a probability distribution, predicted from object and attribute features. As all human descriptions were semantically more insightful than corresponding machine descriptions, we refer the machine descriptions as "sparse" and human descriptions as "rich". Table 2 shows that the distinction between rich and sparse is also evident in the vocabulary and lexical diversity of the datasets.

Figure 1: Rich-Sparse Dataset

3.2 Evaluation Measures on Human and Machine Descriptions

We evaluated human and machine descriptions separately, using the standard evaluation measures. For each image we performed 1000 evaluations, where 5 sentences were randomly selected from the set of descriptions to be the ground truth candidates, with the remaining used to calculate the metrics. Table 1 shows that when both ground truth and candidate description sentences are semantically sparse they perform very well. However descriptions of a higher semantic complexity are penalised as a result of their more diverse and rich descriptions, with many insightful descriptions scoring zero. Figure 3 shows examples where the SPICE metric scores rich descriptions as zero. When rich descriptions were used as ground truth, the machine descriptions perform very poorly.

3.3 Comparison of Lexical Diversity

We measured and compared the LD of human and machine descriptions. Our human descriptions were universally richer and more semantically detailed than the machine descriptions. For each of the 40 TTR

This is a picture of three young men carrying a coffin
The faces of two of the boys who can be seen look serious
Some boys or young men are in a funeral procession

A man is in a hat and hat and tie
A man is talking on a cell phone
Man wearing sunglasses and a red hat with a hat on
A men wearing a red hat and sunglasses

A married couple are celebrating an anniversary
The wome is posing as if she is going to cut the cake
A man and a women are standing next to each other in front of a white cake

Three men are standing around a table with a cake on it
A man cutting a cake on a white table
Three men are standing around a table with a cake on it

A group of Asian Students are preparing to do a class test
They are each reading from a paper which may be a test or an exam
A group exercise which requires students to talk to each other
A group of girls are busy with their classroom activity

A group of kids sitting on a table with laptops
A group of people are sitting around a table
Several people are sitting around a table eating a meal

A young woman is sitting on the floor with a her back against a table leg
A woman is looking sorrowful
The floor is woodern and there are some wooden tables and chairs in the room

The woman is holding a baby and a young boy sitting on the floor
A woman sitting on a wooden bench next to a woman
A little girl sitting on a wooden bench

Figure 2: Examples of human (black) and machine descriptions (red).

A political march is taking place
A protest is taking place
A man is speaking to the marchers through a police traffic cone
A protest scene on a tree lined city street
One protester is using a traffic cone as a megaphone
One woman is holding two flip-flops up in the air
People are shouting and chanting
People are waving their hands in the air and shouting
A man shouting has is holding his arms out and shouting

A crowd of people are standing around a large crowd.
A group of people on a city street.
A group of people on the street with umbrellas
A group of people standing next to a crowd of people.
Many people are walking on a sidewalk with a crowd of people watching.

Figure 3: Zero scoring rich descriptions (top) and low scoring machine descriptions (bottom) when measured on SPICE

curves we plotted (machine and human for each image), we found that LD was an accurate indication of whether a descriptions was from the rich or sparse set. Figure 4 shows the TTR curves for the examples presented in Figure 2 . The figure illustrates the faster decline of the sparse descriptions, relative to the semantically richer descriptions.

Ground Truth	Human		Machine	
Candidates	Human	Machine	Human	Machine
Cider	0.09	0.02	0.01	0.27
Bleu1	0.49	0.37	0.25	0.75
Bleu2	0.22	0.1	0.06	0.59
Bleu3	0.09	0.01	0.01	0.42
Bleu4	0.05	0.00	0.00	0.28
Rouge(L)	0.32	0.23	0.19	0.19
METEOR	0.17	0.1	0.09	0.3
SPICE	0.1	0.05	0.03	0.2

Table 1: Evaluation Measures for Rich (Human) and Sparse(Machine) Domains

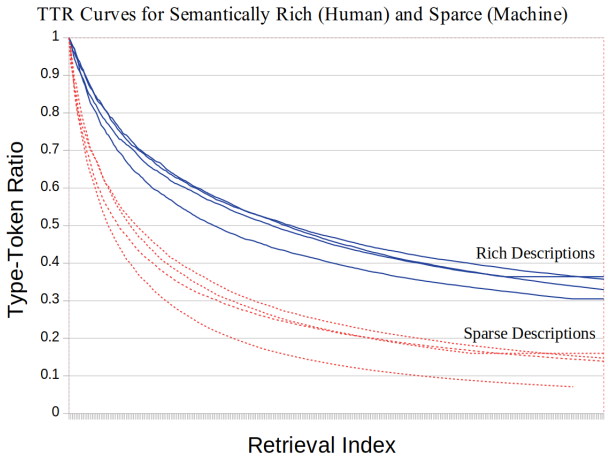

Figure 4: TTR curves for rich and sparse descriptions

3.4 Comparison of Linguistic Complexity

Readability measures have long been used to automatically grade the complexity of language. We tested several measures, including Flesch–Kincaid(Kincaid JP, 1988), Coleman–Liau(Coleman and Liau, 1975), Dale-Chall(Dale E, 1948) and Automated Readability(Senter, 1967). However we found they did not correlate well with semantic quality. Informative descriptions tend to be lexically diverse, but are not necessarily complex. Rich descriptions can contain a higher syllable count and more 'difficult' words than sparse descriptions however this is not always the case. Furthermore a description corpus which generates exactly the same complex sentence for every image conveys no information and yet would score highly on complexity.

4 The Lexical Gap

One indication of the performance of a machine description system, is its ability to convey semantically rich information. We propose a measure which considers the entire output of a description system (which

	Lexical Diversity				
	TTR	Root-TTR	Log-TTR	HDD	MTLD
Sparse	0.09	2.98	0.65	0.55	16.07
Rich	0.24	14.29	0.83	0.75	40.58

Table 2: Rich-Sparse Dataset Statistics

we call c_m) and compares it with its training data (which we call c_r). Thus instead of solely considering a machine's ability to predict n-grams or words, we also measure its ability to maintain the linguistic diversity of its training corpus. Our key finding is that measuring the LD of a description corpus relative to its ground truth data is a good indication of semantic quality, and can be used to weight standard performance measures, increasing their correlation with human subjective judgement. In this section we define our measures, which we later compare compare with standard captioning measures.

4.1 Measuring the Lexical Gap

The Lexical Diversity Ratio (LDR) is a straightforward measure of the ability of a machine to match the semantic depth of its source material. Given a function L which calculates LD for a reference description corpus c_r and the machine description corpus c_m, we define the Lexical Diversity Ratio (LDR) l_d as:

$$l_d = \frac{L(c_m)}{L(c_r)} \tag{1}$$

A machine with a score of 1, is more able to match the lexical diversity of its training source. A lower score, indicates a reduction in semantic richness. We also define the lexical gap (L_g) a bounded measure of the ability of a system to maintain lexical diversity. An l_d below some constant μ, will tend to zero indicating a larger lexical gap. As l_d increases a system is closing that gap, towards a score of 1, which indicates ideal performance. Given the constants μ and α, we define the Lexical Gap L_g:

$$L_g = \frac{1}{1 + \exp^{-\alpha(l_d - \mu)}} \tag{2}$$

Considering our rich and sparse descriptions independently, we split them into sub-corpora. We calculate l_d scores each sub-corpora as (c_r) using in every case the richer descriptions has our reference c_r. Figure 5 shows the LDRs (l_d) for the rich and sparse parts of our description dataset. The richer descriptions, although more broadly distributed, have a higher mean l_d. We define μ as the value that produces the Bayes Minimum error between the two distributions of l_d (0.81), and we set α=5 to distribute all our values broadly and between the range 0..1. Then given a description metric M, we calculate the gap-

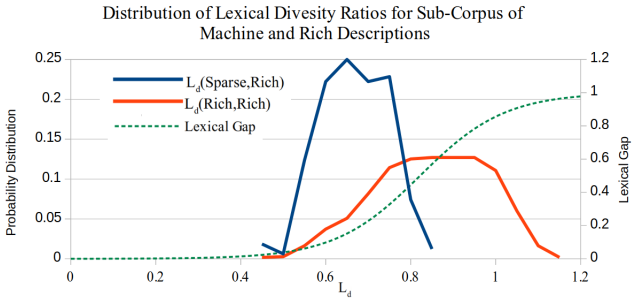

Figure 5: Distribution of LDR scores for sparse and rich descriptions

weighted score for each sentence: s_n in a corpus $s_n \subseteq c_r$:

$$m_{gap} = M(s_n)L_g \tag{3}$$

$$m_{ldr} = M(s_n)L_d \tag{4}$$

5 Results

We evaluated the performance of weighted lexical measures using the Composite dataset. The dataset contains selected human and machine descriptions for images sourced from Flickr30k, Flickr8K and

Source Dataset	Caption Source	LDR (l_d)	Lexical Gap (L_g)
Flickr30k	Human	1.03	0.98
	Machine1	0.63	0.02
	Machine2	0.70	0.08
	Machine3	0.71	0.11
Flickr8k	Human	0.92	0.89
	Machine1	0.71	0.10
	Machine2	0.65	0.03
MS COCO	Human	0.97	0.95
	Machine1	0.72	0.11
	Machine2	0.73	0.13
	Machine3	0.73	0.13

Table 3: Calculation of l_d and L_g for the Composite Dataset

	Spearman	Pearson	Kendal-T
NNEval	**0.524**	0.532	**0.404**
l_d	0.473	0.621	0.329
L_g	0.473	**0.630**	0.369

Table 4: Overall Correlations for LDR and Lexical Gap

MS COCO. For each description in Composite, we sourced the relevant ground truth sentences from the source dataset so that we could calculate the captioning scores for that sentence. These are the standard scores presented in Table 5.

Using our measures defined previously, we also calculated l_d and L_g for each subset of the Composite dataset (Table 3) using the relevant source corpus as our reference (c_r). We thus measured the lexical diversity of human and machine subsets of the Composite dataset. Before using standard evaluation measures, we found that our l_d and L_g correlated well with human subjective judgements, as presented in Table 4. Then we calculated the m_{gap} and m_{ldr} for each evaluation measure over the entire Composite dataset. We calculate the correlation performance with the human evaluation scores.

Table 5 compares the gap weighted scores with standard measures of performance. We found that on all measures, weighting by l_d and L_g improves the correlation between human judgements and objective measures.

	Spearman			Pearson			Kendal-T		
	Standard	m_{ldr}	m_{gap}	Standard	m_{ldr}	m_{gap}	Standard	m_{ldr}	m_{gap}
CIDEr	0.361	0.383	**0.516**	0.354	0.388	**0.571**	0.270	0.369	**0.389**
Bleu1	0.346	0.429	**0.444**	0.362	0.471	**0.489**	0.257	0.292	**0.362**
Bleu2	0.323	**0.395**	0.393	0.342	0.411	**0.534**	0.258	**0.283**	0.282
Bleu3	0.292	0.382	**0.516**	0.286	0.327	**0.544**	0.250	0.277	**0.392**
Bleu4	0.235	0.373	**0.531**	0.202	0.228	**0.569**	0.206	0.286	**0.401**
Rouge_L	0.364	0.447	**0.473**	0.369	0.476	**0.632**	0.271	0.319	**0.369**
Meteor	0.367	0.427	**0.473**	0.400	0.478	**0.635**	0.275	0.335	**0.369**
SPICE	0.372	0.409	**0.540**	0.399	0.448	**0.573**	0.299	0.329	**0.411**

Table 5: Overall Correlations for LDR and Lexical Gap. All p-values$<$0.001

6 Conclusion

Much progress has been in visual description, with many systems capable of generating original sentences which convey salient objects and attributes. However building systems capable of conveying semantically insightful information still remains a big challenge because of the difficulty of developing effective and insightful evaluation measures. We find that LD of descriptions is a useful indicator of semantic quality, and propose that description systems are measured not only on the accuracy of their predictions, but also on their ability convey lexically specific information. Measuring LD, rewards systems which are able to preserve rich and diverse descriptions, but penalises sparse systems, which have a poor lexical capability.

We hope that our work will inspire larger datasets of semantically richer and more detailed descriptions, and the development of more effective evaluation criteria for descriptions.

References

Aafaq, N., S. Z. Gilani, W. Liu, and A. Mian (2018). Video Description: A Survey of Methods, Datasets and Evaluation Metrics. pp. 1–25.

Aditya, S., Y. Ang, C. Baral, C. Fermuller, and Y. Aloimonos (2015). From Images to Sentences through Scene Description Graphs using Reasoning and Knowledge. *Arxiv*.

Anderson, P., B. Fernando, M. Johnson, and S. Gould (2016). SPICE: Semantic propositional image caption evaluation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 9909 LNCS, pp. 382–398.

Anderson, P., X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang (2017). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.

Bernardi, R., R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research 55*, 409–442.

Chen, D. L. and W. B. Dolan (2011). Collecting Highly Parallel Data for Paraphrase Evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Coleman, M. and T. L. Liau (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology Vol. 60*, pp. 283–284.

Dale E, C. J. (1948). A Formula for Predicting Readability. *Educational Research Bulletin 27*(2), 37–54.

Denkowski, M. and A. Lavie (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *WMT*, pp. 376–380.

Durán, P., D. Malvern, B. Richards, and N. Chipere (2004). Developmental trends in lexical diversity.

Hodosh, M., P. Young, and J. Hockenmaier (2015). Framing image description as a ranking task: Data, models and evaluation metrics. In *IJCAI International Joint Conference on Artificial Intelligence*, Volume 2015-Janua, pp. 4188–4192.

Jain, U., S. Lazebnik, and A. Schwing (2018). Two can play this Game: Visual Dialog with Discriminative Question Generation and Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5754–5763.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning : A journal of Research in Language 63*(1), 87–106.

Kincaid JP, Braby R, M. J. (1988). Electronic authoring and delivery of technical information. *Journal of Instructional Development*.

Lin, C.-Y. and E. Hovy (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03 2003*(June), 71–78.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 8693 LNCS, pp. 740–755.

Long, J., E. Shelhamer, O. Vinyals, A. Toshev, S. Bengio, D. Erhan, K. Lenc, A. Vedaldi, E. Denton, S. Chintala, A. Szlam, R. Fergus, P. Fischer, H. Philip, C. Hazırbas, P. V. D. Smagt, D. Cremers, T. Brox, F. Meng, Z. Lu, Z. Tu, H. Li, Q. Liu, V. Mahadevan, and S. Member (2014). Show and Tell: A Neural Image Caption Generator. *arXiv 32*(1), 1–10.

McCarthy, P. M. and S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (July), 311–318.

Plummer, B. A., L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik (2017). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*.

Schuster, S., R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning (2015). Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. *Emnlp*, 70–80.

Senter, R.J.; Smith, E. (1967). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report.

Sharif, N., L. White, M. Bennamoun, and S. A. A. Shah (2018). NNEval: Neural network based evaluation metric for image captioning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 11212 LNCS.

Teney, D., L. Liu, and A. v. d. Hengel (2016). Graph-Structured Representations for Visual Question Answering. pp. 1–9.

Vedantam, R., C. L. Zitnick, and D. Parikh (2015). CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 07-12-June, pp. 4566–4575.

Wu, Q., C. Shen, A. v. d. Hengel, P. Wang, and A. Dick (2016). Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*(6), 1367–1381.

Xu, K., J. L. B. R. Kiros, K. C. A. Courville, and R. S. R. S. Z. Y. Bengio (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *IEEE Transactions on Neural Networks 5*(2), 157–166.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, Volume 2015 Inter, pp. 19–27.