Exploring the Impact of Training Data Bias on Automatic Generation of Video Captions

Alan F. Smeaton(✉), Yvette Graham, Kevin McGuinness, Noel E. O'Connor, Seán Quinn, and Eric Arazo Sanchez

Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland
alan.smeaton@dcu.ie

Abstract. A major issue in machine learning is availability of training data. While this historically referred to the availability of a sufficient volume of training data, recently this has shifted to the availability of sufficient unbiased training data. In this paper we focus on the effect of training data bias on an emerging multimedia application, the automatic captioning of short video clips. We use subsets of the same training data to generate different models for video captioning using the same machine learning technique and we evaluate the performances of different training data subsets using a well-known video caption benchmark, TRECVid. We train using the MSR-VTT video-caption pairs and we prune this to reduce and make the set of captions describing a video more homogeneously similar, or more diverse, or we prune randomly. We then assess the effectiveness of caption-generating trained with these variations using automatic metrics as well as direct assessment by human assessors. Our findings are preliminary and show that randomly pruning captions from the training data yields the worst performance and that pruning to make the data more homogeneous, or diverse, does improve performance slightly when compared to random. Our work points to the need for more training data, both more video clips but, more importantly, more captions for those videos.

Keywords: Video-to-language · Video captioning
Video understanding · Semantic similarity

1 Introduction

Machine learning has now become the foundation which supports most kinds of automatic multimedia analysis and description. It is premised on using large-enough collections of training data, which might be labelled images, or captioned videos, or spoken audio with text transcriptions. This training data is used to train models of the analysis or description process and these models are used to analyse new and unseen multimedia data, thus automating the process.

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11295, pp. 178–190, 2019.
https://doi.org/10.1007/978-3-030-05710-7_15

Machine learning algorithms used to train the automatic process are improving with a current concentration on deep learning and a recent focus on multimodal learning, i.e. learning from multiple sources [5]. However, there is a veritable zoo of algorithmic techniques and possible approaches available [11] as well as a constant stream of new emerging ideas. It is reasonable to say that choosing the best machine learning algorithm from those available requires significant prior knowledge making the process akin to a "black art" with little underlying understanding of why different approaches work better in different applications, let alone a unified theory. Another issue is training data. While this used to refer to the availability of a sufficient *volume* of training data, recently this has shifted to the availability of sufficient *unbiased* training data, or rather an awareness of existing biases within training data.

In this paper we focus on the effect of training data bias on an emerging multimedia application, the automatic captioning of short video clips. We use variations of the same training data to generate different models for video captioning using the same machine learning techniques and we evaluate the performance of different training sets using the TRECVid benchmark.

This paper is organised as follows. Section 2 introduces related work covering data bias and training data, techniques used for automatic video captioning, and related work focused on data bias in training data for video captioning. Section 3 describes our experimental setup, training and test data used, the captioning models selected, and the metrics to assess caption quality. Section 4 presents our experimental results, and an analysis of those results is in the concluding section.

2 Related Work

2.1 Data Bias and Training Data

Bias exists in all elements of society, and in almost all data we have gathered. These biases are both latent and overt and influence the things we see, hear and do [4]. Biases are an intrinsic part of our society, and always have been, but so long as we are aware of such biases we can compensate and allow for them when we make decisions based on such biased data.

The data gathered around our online activities, our interactions with the web and its content and our interactions through social media is particularly prone to including biases. This is because much of our online activity is self re-enforcing, building on similarity and overlap by, for example, recommending products or services which we are likely to use because we use similar ones, or forming social groups where homogeneity rather than diversity is the norm.

Different forms of bias exist in our online data covering social and cultural aspects like gender, age, race, social standing, ethnicity as well as algorithmic bias covering aspects like sampling and presentation. While this may be undesirable as a general point, it becomes particularly problematic when we then use data derived from our online interactions, as a driver for some algorithmic process. In [4] the author points out that recommending labels or tags for images or videos is an extreme example of algorithmic bias when it is based on similarity with

already tagged images or videos and/or used in collaborative filtering. In such a case, which is widespread, there is no novelty, no diversity, no and enlarging of the tag set, and this is a point we shall return to later.

2.2 Automatic Video Captioning: Video-to-text

Automatic video captioning or video description is a task whereby a natural language description of a video clip is generated which describes video content in some way. It is a natural evolution of the task of automatic image or video *tagging* which has seen huge improvement within the last few years to the extent that automatic techniques now replace manual tagging on a web-scale.

Video description or captioning has many useful applications including uses in robotics, assistive technologies, search and summarisation, and more. But video, or even image, description is extremely difficult because images and videos are so information-rich that reducing their content to a single caption or sentence is always going to fall short of capturing the original content with all its nuances. Issues of vocabulary usage, interpretation, bias from our background culture or current task or context, all contribute to it being almost impossible to get a universally agreed caption or description for a given video or image.

Broadly speaking, there are three approaches to automatic image and video captioning. The first, called *pipelined*, aims to recognise specific objects and actions in the images/videos and uses a generative model to create captions. This has the advantage that it builds on object/action detection and recognition whose quality is now quite good, and it can generate new captions not seen in the training data. The second approach involves projecting captions and videos from a training set into a common representation or space and *finding the closest existing video* to the target and using its caption. The disadvantage of this approach is that it cannot *create* new captions, it just re-uses existing ones. The third approach, which has become popular, is an *end-to-end solution*. It uses a pre-trained CNN such as VGG16, Inception or ResNet, to extract a representation and using this as input to a RNN-based caption generator. This approach can generate novel captions but it requires plentiful training data.

A good description of recent work in video description, including available training material and evaluation metrics, can be found in [1]. While this is a good survey, there is even more recent work appearing in the literature such as the convolutional image captioning work in [2], where they do away with the LSTM decoder altogether and show that you can generate good captions just using temporal convolutions. Other recent work focuses on dense video captioning, captioning longer videos with many captions, aiming to generate text descriptions for all events in an untrimmed video [21] as opposed to other work which sets out to caption short video clips of just a few seconds duration.

2.3 Data Bias in Video Captioning

Given the recognised existence of bias in almost all our data, and the dependence on training data for training machine learning algorithms, it is inevitable that

biases will affect the performance of video captioning systems. This is especially so considering the open nature of the domain of video captioning where almost anything can be captioned, yet very little work has been reported on assessing the impact of data bias on the quality of generated captions.

Much of the work which tries to map video clips to natural language and vice-versa, sets out to map both language (text) and images (or videos) to a common space, such as in [16]. When working with such an approach it is important that the multimedia artifacts, whether images/videos or text fragments (captions) mapped into a common space are distinct and that there is "distance" between them. Separating the artifacts is essential to allow whatever kind of content-based operation is being developed. This also reveals that biases in the data, among any sets of objects, will yield clusters of similar objects in the common space and this will be unhelpful and so achieving diversity among the objects in the common space is important, a point highlighted in [12].

More recent work reported in [17] highlighted the difficulties in evaluating the quality of multiple captions for the same video. In an attempt to achieve diversity and coherence in training data in their work, the authors removed outlier captions from the MSR-VTT training dataset by using SenseEmbed, a method to obtain a representation of individual word *senses* [10] which allowed them to model different senses of polysemous words. Outlier captions are those whose semantic similarity among the set of captions for a single video clip make them different from the rest of the captions, or far removed from a centroid. Individual word sense embeddings are combined to create a global similarity between captions that is used to determine outliers, which are then pruned. The authors also added new captions that mix-and-match among subject, object and predicate in the original captions. Using the MSR-VTT training data [20], results indicate small improvements in generated captions when outlier captions are removed from the training data.

In this paper we take the work in [17] further by judiciously removing some captions from the training data, not just because they may be outliers but because they may help improve overall diversity or homogeneity of the training data.

3 Experimental Setup

Figure 1 outlines our experimental procedure. We start with a collection of 10,000 short videos (1) from the MSR-VTT collection [20]. Each has been manually captioned 20 times with a sentence descriptor (2) some of which may be duplicates. For each video, we computed inter-caption semantic similarity (190 pair-wise caption similarity computations per video) using the STS measure of semantic similarity described in [9,13]. STS similarities is based on distributional similarity and Latent Semantic Analysis (LSA) complemented with semantic relations extracted from WordNet. We use the STS similarity values to prune captions from each set of 20 captions per video.

Our first approach to pruning captions is to randomly remove them, reducing the 200,000 to about 160,000 overall, shown as step (5) in the diagram with

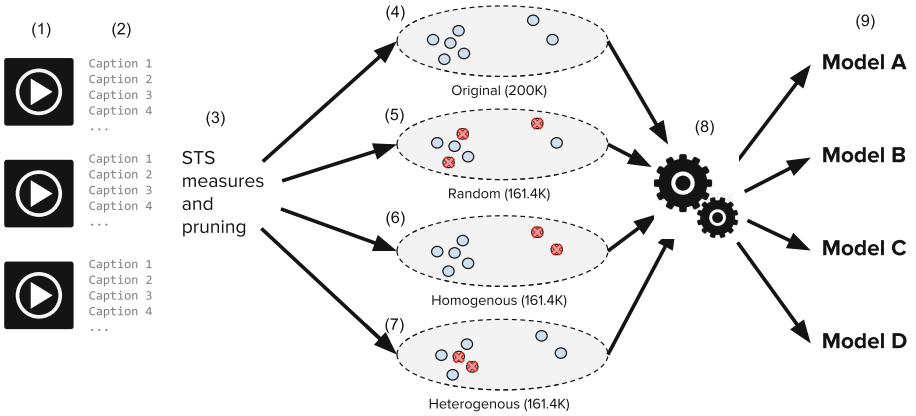


Fig. 1. Outline of experimental setup (Color figure online)

the red crosses indicating captions which are removed. A second approach we take is to remove captions which are semantically distinct from the set of other captions, resulting in a more homogeneous set of captions with stronger overall inter-caption similarity. We refer to this strategy as "homogeneous", shown as step (6). The inverse of this is to remove captions that are already semantically similar to other captions in the set of captions for a single video, an approach we refer to as generating a more "heterogeneous" collection of captions (step 7).

We use the four variations of the MSR-VTT training data [20], derived from the full collection and described in Sect. 4.1 – the full collection, randomly pruned collection, homogeneously pruned collection and heterogeneously pruned collection – to each train an individual model for caption generation (8), which in turn generates 4 models referred to as A, B, C, and D.

3.1 Training Data for Video Captions

There is already a host of training data available for generating video captions and currently the best source of information on this is in [1]. This presents details of publicly-available training datasets, covering MSVD, MPII Cooking, YouCook, TACoS, TACos-MLevel, MPII-MD, M-VAD, MSR-VTT, Charades, VTW and ActyNet Cap. These vary from 20,000 videos of 849 h duration (ActyNet CAP) to just 2.3 h (YouCook).

The data we use in this paper is from the Microsoft Research Video to Text (MSR-VTT) challenge [20] a large-scale video benchmark. The videos are of 41.2 h duration in total, each clip annotated with 20 natural sentences by 1,327 AMT workers forming the groundtruth for captions.

3.2 Video Caption Generation

To evaluate how the training data variations affect video captioning, we used an end-to-end stable model that generates natural language descriptions of short

video clips from training data. We selected the S2VT model from [19] that consists of a stack of two LSTMs, one to encode the frames and another that uses the output of the first LSTM to generate a natural language caption. Both LSTMs have 1,000 hidden units. This is a sequence-to-sequence model that generates a variable length sequence that corresponds to the caption, given a video with a variable number of frames.

Video frames are encoded using VGG16 pre-trained on ImageNet with weights fixed during training. The resulting 4096D representation from *fc7* (after the ReLU) are projected to 500D for input to the first LSTM, which encodes multiple frame representations into a single representation that captures the visual and temporal information from the clip. During this stage the output of the second LSTM is ignored. The output of the first LSTM is passed to the second LSTM together with a "beginning of sentence" tag to generate the first word of the caption. Subsequent words are generated by concatenating the previous predictions to the output of the first LSTM until the"end of sentence" tag is predicted.

To obtain the words we project the output of the second LSTM to a 22,939D vector. We built the vocabulary from the captions in MSR-VTT [20] without any pre-processing of the words. We then apply a softmax to find the predicted word in each step. The model is trained to minimise the cross entropy between the predicted words and the expected output using SGD on the MSR-VTT training set. We trained for 25,000 iterations, where an iteration consists of a batch of 32 videos. We sub-sample 1 in 10 frames to accelerate computation.

3.3 Evaluating Quality of Automatic Video Captions

Evaluating the performance of automatic captioning in a way that balances accuracy, reliability and reproducibility is a challenge that may be irreconcilable. Current approaches assume there is a collection of video clips with a reference or gold standard caption against which to compare, yet we know there can be many ways to describe a video so who is to know what is and is not correct? As a default, most researchers work with a reference caption and use measures including BLEU, METEOR, or CIDEr to compare automatic vs manual captions but this is an active area of research and there are no universally agreed metrics.

BLEU has been used in machine translation to evaluate the quality of generated text. It approximates human judgement and measures the fraction of N-grams (up to 4-gram, so BLEU1, BLEU2, BLEU3 and BLEU4) in common between target text and a human-produced reference. BLEU's disadvantage is that it operates at a corpus level and is less reliable for comparisons among short, single-sentence captions. METEOR computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. It is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision and like BLEU it also operates at a corpus level rather than at a set of independent video captions. CIDEr computes the TF-IDF (term frequency inverse document

frequency) for each n-gram of length 1 to 4 and is shown to agree well with human judgement.

A number of benchmarks have emerged in recent years to assess and compare approaches to video caption generation and one of those is the VTT track in TRECVid in 2016 and 2017 [18]. In the 2017 edition [3], 13 groups participated and submitted sets of results including ourselves [15] and we re-used the infrastructure from that in the work reported here.

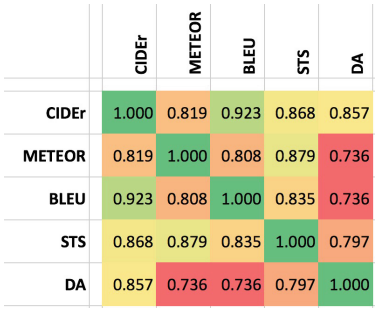
In the TRECVid 2017 VTT task, the STS measure [9], mentioned earlier and used by in this work to determine semantic outlier captions among the 20 manual captions assigned to each video in the MSR-VTT collection, was used to measure the similarity between captions submitted by participants and a manual reference. We include mention of STS as an evaluation measure but do not use this as a measure; however, we do use it to compute similarity among manual captions for the same video.

The final evaluation metric is known as direct assessment (DA) and it brings human assessment using Amazon Mechanical Turk (AMT) into the evaluation by crowdsourcing how well a given caption describes its corresponding video. The metric is described in [7] and includes a mechanism for quality control of the ratings provided by AMT workers via automatic degradation of the quality of some manual captions hidden within AMT HITs (Human Intelligence Tasks). In this way, DA produces a rating of the reliability of each human assessor and filters out any unreliable human assessors prior to producing evaluation results. It provides a reliable way to distinguish genuine human assessment from attempts to game the system by augmenting the training data similar to what was done in [17] but not to increase the amount of training data but to validate the accuracy of the AMT workers. A human assessor is required to rate each caption on a [0..100] rating scale and the ratings are micro-averaged per caption before computing the overall average for the system (called RAW DA). The average DA score per system is also computed after standardisation per individual crowdsourced worker's mean and standard deviation score (called Z-score).

In a recent analysis of metrics for measuring the quality of image captions [14], the authors introduced a variation of Word Mover's Distance (WMD) and compared this against the other "standard" metrics but not direct assessment, concluding that they are each significantly different to each other. Work described in [1] also examined different evaluation metrics and concluded that evaluation is more reliable when more reference captions are available to compare against, and that CIDEr and METEOR seem to work best in such situations.

To test this we took system rankings from 13 participants in the TRECVid 2017 VTT task according to the CIDEr, METEOR, BLEU, STS and DA metrics presented in [3] and calculated Spearman's correlation among pairs of rankings. The results (Fig. 2) show good agreement among the automatic metrics (CIDEr, METEOR and BLEU) with STS and DA being somewhat different both from each other and from the automatic systems, but with a lowest correlation of 0.736 there is still reasonable agreement among all metrics. What all this means

in terms of evaluation metrics for this work is that we should consider all metrics when assessing the performance of a video caption generation system.



	CIDEr	METEOR	BLEU	STS	DA
CIDEr	1.000	0.819	0.923	0.868	0.857
METEOR	0.819	1.000	0.808	0.879	0.736
BLEU	0.923	0.808	1.000	0.835	0.736
STS	0.868	0.879	0.835	1.000	0.797
DA	0.857	0.736	0.736	0.797	1.000

Fig. 2. Metric correlations for evaluating video caption system performance in [3].

4 Experimental Results

4.1 Pruning the Training Data

Two strategies were developed to prune captions from our training data based on the semantic text similarity (ground) measure described in [13]. This method takes two segments of text and returns a score in the range [0..1], representing how similar the pieces of text are in their semantic meaning. Our training set contained 10,000 videos with 20 human captions per video. To implement our pruning of captions, we first calculated inter-caption STS scores for all caption pairings on each video. This resulted in 190 inter-caption similarity scores per video. We denote the inter-caption similarity between a caption A and caption B as as $sim(A, B)$. We then computed average similarity scores for each caption by averaging the 19 inter-caption scores for that caption. We denote the average inter-caption similarity for a caption A as $avgsim(A)$. We computed summary statistics on the entire populations of $avgsim$ and $sim(A, B)$ scores to allow us to establish appropriate thresholds for our pruning as outlined in the following.

Homogeneous Pruned Training Dataset. The homogeneous pruning strategy aimed to remove captions which were highly dissimilar to the other captions for a given video, thus removing the "outlier" captions in our training data. There are two requirements which must be met for caption X to be pruned under this strategy:

1. $avgsim(X)$ must be below the 30th percentile for our total population of $avgsim$ scorings. For the dataset used in this experiment the 30th percentile $avgsim$ threshold was 0.36099.

2. There must not exist any caption Y for the same video as X where $sim(X, Y)$ is greater than the 80th percentile for our total population of $sim(A, B)$ scorings. For the dataset used in this experiment the 80th percentile $sim(A, B)$ threshold was 0.610.

Requirement 1 asserts that caption X can be considered dissimilar to all the other captions for the video, while requirement 2 asserts that there is no other caption Y present for this video which is highly similar to this caption X. This strategy resulted in the pruning of 38,379 captions.

Heterogeneous Pruned Training Dataset. The heterogeneous pruning strategy aimed to reduce the size of clusters of captions which were highly similar to each other, thus enforcing greater diversity in our training data. Similar to the threshold used in our homogeneous pruning we define two captions X and Y to be highly similar or "neighbours" if $sim(X, Y)$ is greater than the 80th percentile for our total population of $sim(A, B)$ scorings, which for our dataset is 0.610. The procedure for heterogeneous pruning is as follows:

1. For each video rank the captions using the number of neighbours they have.
2. If the highest neighbour count is greater than 3, prune this caption and recalculate neighbour counts.
3. Continue pruning the caption with highest neighbour count until no caption has more than 3 neighbours.

This resulted in the pruning of 38,816 captions.

Random Pruned Training Dataset. Here we randomly chose 38,598 captions to be pruned from across the 10,000 videos. This is the average of the number pruned by heterogeneous and homogeneous strategies.

In terms of overall changes to the data as a result of pruning we did not compute whether the overall vocabulary has reduced and if so by how much. In terms of data availability, the original MSR-VTT-10k dataset is openly available and our prunings of this according to homogeneous, heterogeneous and random strategies, is available at https://github.com/squinn95/MMM_2018_Files/.

4.2 Performance Figures for Generating Video Captions

Tables 1 and 2 show results using the automatic metrics (BLEU1, 2, 3, 4, METEOR, and CIDEr) computed using the code released with the Microsoft COCO Evaluation Server [6] and the direct assessment (DA) evaluation for both the MSR-VTT and the TRECVid 2017 collections. When computing DA, we also compute DA scores for the manual (human) captions for both collections as reference points. In the case of TRECVid 2017 (Table 2) we reproduce the official DA values for the TRECVid assessment of human captions as well as the best-performing of the official submissions.

Table 1. Results for MSR-VTT17 videos using 4 different MSR-VTT training datasets

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	CIDEr	DA	
							Av	z
HOM	75.0	57.0	40.8	27.6	20.7	18.6	60.2	−0.066
DIV	73.7	56.2	40.2	26.8	20.2	16.3	59.1	−0.090
RAND	75.1	57.0	40.6	27.2	20.9	18.9	58.8	−0.110
ALL	73.5	57.2	41.9	29.2	20.7	18.0	57.8	−0.131
Human captions							88.6	0.690

Table 2. Results for TRECVID17 videos trained using 4 different MSR-VTT training datasets, manual (human) annotation evaluation and manual (human) and best automatic results from TRECVid 2017[3]

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	CIDEr	DA	
							Av	z
HOM	24.3	11.4	5.5	2.8	8.4	15.6	49.7	−0.151
DIV	22.3	9.9	4.6	2.1	8.1	13.8	48.5	−0.180
RAND	24.6	11.0	5.2	2.3	8.3	14.3	47.2	−0.205
ALL	24.5	11.1	5.2	2.5	8.6	16.1	50.1	−0.150
Human captions							82.8	0.723
TRECVid 2017 Human captions							87.1	0.782
TRECVid 2017 Best automatic performance (RUC_CMU)							62.2	0.119

In terms of human evaluation, raw DA scores for all runs range from 57.8 to 60.2% for the MSR-VTT dataset, with HOMogeneous training achieving highest absolute DA score, and from 47.2 to 50.1% for the TRECVid dataset, with training on ALL data achieving the highest DA score overall.

DA scores achieved by competing runs are close for all runs and tests for statistical significance should be carried out before concluding that differences in performance are not likely to occur simply by chance. We therefore carry out Wilcoxon rank-sum test for DA scores for both sets of test data. In the case of both datasets, as expected, all runs were significantly lower than ratings achieved by human captions. Competing runs also showed no significance difference in performance with the single exception of ALL on the TRECVid test set achieving a significantly higher DA score compared to that of RAND.

In contrast in terms of metric scores, competing runs showed mixed results in terms of ordering of performances. For the TRECVid-tested data collection BLEU indicates best performance for RAND and HOM, while ALL achieves best performance according to Meteor and CIDEr. On the MSR-VTT dataset, it is ALL that achieves best performance according to Meteor and CIDEr while BLEU indicates top performance for RAND. Although testing for statistical significant differences in BLEU and other metric scores is common in Machine Translation evaluation using Bootstrap resampling or Approximate Randomization for example [8], we do not report these here as the accuracy of such methods has not as yet been tested for the purpose of video captioning.

The best-performing automatic submission for the DA metric in TRECVid on TRECVid17 data was from Renmin University of China and Carnegie Mellon University (RUC_CMU) (Table 2) with DA values of Av = 62.2, z = −0.119. This is higher than the highest of our automatic systems, which had Av = 50.1, z= −0.150. While we would have liked to work with a better performing caption generation system and ours is not as good as the best possible but it was sufficient to allow us to experiment with different sets of training data and evaluate their effectiveness.

The DA values for our assessment of human captions in TRECVid 2017 are lower than the official TRECVid assessment of the same (Av 82.8 vs 87.1) but variance in DA scores from different sets of Mechanical Turk workers are to be expected. Additionally, there is a pool of human captions for each video in this dataset and human captions were chosen at random for each evaluation. Either way, as expected, human captions are significantly better than all other runs in each dataset, showing that automatic systems still need improvement.

5 Analysis and Conclusions

We expected that improving diversity in training data would achieve better performance in the resulting caption generation, as advocated in [12] as opposed to other work to improve the quality of training data which simply removed outliers in [17] for example. In practice we did not find this and like the work in [17] the improvements are minor for pruned datasets. Based only on statistical significance, all system variations are roughly the same performance with the exception that RAND performs significantly worse than ALL on the TRECVid dataset. It is possible that given a larger training set this difference (RAND and ALL) might increase and distinguish RAND from all other runs, but it could just as easily be that given a larger test set the difference disappears. Also, the biases we address here are based on semantic similarity between captions which is a limited form of bias and doesn't address biases as a result of gender, ethnicity, etc. which will require digging deeper into the semantics of homogeneity and diversity criteria.

With most of the work in machine learning applications there is a seemingly insatiable desire for more and better training data, and in our case this means more video clips, and increased volumes of manual captions for each clip. The availability of video clips is not a problem but we are unlikely to realise increases in manual captioning of these videos with the approaches used to date. Instead we could look at pre-processing existing manual captions to generate variations for the videos we already have using data augmentation techniques like synonym substitution and others used as part of the STS measure. This forms part of our planned future work.

Acknowledgements. This work is supported by Science Foundation Ireland under grant numbers 12/RC/2289 and 15/SIRG/3283.

References

1. Aafaq, N., Gilani, S.Z., Liu, W., Mian, A.: Video description: a survey of methods, datasets and evaluation metrics. arXiv preprint arXiv:1806.00186 (2018)
2. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: Computer Vision and Pattern Recognition (CVPR), June 2018
3. Awad, G., et al.: TRECVID 2017: evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In: Proceedings of TRECVID 2017. NIST (2017)
4. Baeza-Yates, R.: Bias on the web. Commun. ACM **61**(6), 54–61 (2018)
5. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. (Early Access) (2018). https://doi.org/10.1109/TPAMI.2018.2798607
6. Chen, X., et al.: Microsoft COCO captions: data collection and evaluation server. CoRR, abs/1504.00325 (2015)
7. Graham, Y., Awad, G., Smeaton, A.: Evaluation of automatic video captioning using direct assessment. CoRR, abs/1710.10586 (2017)
8. Graham, Y., Mathur, N., Baldwin, T.: Randomized significance tests in machine translation. In: ACL 2014 Workshop on Statistical Machine Translation, pp. 266–274. Association for Computational Linguistics (2014)
9. Han, L., Kashyap, A., Finin, T., Mayfield, J., Weese, J.: UMBC EBIQUITY-CORE: semantic textual similarity systems. Joint. Conf. Lex. Comput. Semant. **1**, 44–52 (2013)
10. Iacobacci, I., Pilehvar, M.T., Navigli, R.: SensEmbed: learning sense embeddings for word and relational similarity. In: Proceedings of ACL, pp. 95–105 (2015)
11. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
12. Karpathy, A.: Connecting images and natural language. Ph.D. thesis, Stanford University, August 2016
13. Kashyap, A., et al.: Robust semantic text similarity using LSA, machine learning, and linguistic resources. Lang. Resour. Eval. **50**(1), 125–161 (2016)
14. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E.: Re-evaluating automatic metrics for image captioning. In: Proceedings of EACL, April 2017
15. Marsden, M., et al.: Dublin City University and partners' participation in the INS and VTT tracks at TRECVid 2016. In: Proceedings of TREVid, NIST, Gaithersburg, MD, USA (2016)
16. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: Computer Vision and Pattern Recognition (CVPR), pp. 4594–4602 (2016)
17. Pérez-Mayos, L., Sukno, F.M., Wanner, L.: Improving the quality of video-to-language models by optimizing annotation of the training material. In: Schoeffmann, K., et al. (eds.) MMM 2018. LNCS, vol. 10704, pp. 279–290. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73603-7_23
18. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR 2006: International Workshop on Multimedia Information Retrieval, pp. 321–330 (2006)

19. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: International Conference on Computer Vision (ICCV) (2015)
20. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: Computer Vision and Pattern Recognition (CVPR), pp. 5288–5296, June 2016
21. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Computer Vision and Pattern Recognition (CVPR), June 2018