Specificity-Controlled Video Captioning

Xinyu Liu Ellie Pavlick Daniel Ritchie George Konidaris Stefanie Tellex

Abstract

This paper introduces a specificity-controlled video captioning (SCVC) model that generates one-sentence descriptions of open-domain videos at a level of granularity specified by the user. Previous video captioning models produce captions at inconsistent levels of specificity and cannot explicitly control the amount of detail contained in a caption. In contrast, SCVC produces captions conditioned on both a video and a target specificity level. The SCVC model uses a convolutional neural network (CNN) to first extract visual features from video frames then an attention-based sequence to sequence model (Seq2Seq-Attn) to generate captions. Feature-wise linear modulation (FiLM) layers are inserted in Seq2Seq-Attn to control the specificity level of the generated captions.

We design an automatic evaluate scheme specifically for SCVC tasks to measure the caption and specificity fidelity of generated captions. A comparison between SCVC and ablated models show significant improvement in specificity correctness while maintaining caption correctness. We also qualitatively evaluate the efficacy of the SCVC model and observe desired levels of specificity in the produced captions while maintaining the caption quality of the baseline video captioning model that does not support specificity control.

1. Introduction

Real-world open-domain videos contain hierarchical and compositional structures in the spatial and temporal dimensions. Reasoning about streams of complex visual input and generating coarse- or fine-grained descriptions in a controlled manner lie at the intersection of computer vision and natural language processing. Automatically generating captions of a video at different levels of granularity requires detection and tracking of objects in a dynamic scene and extract compact representations of spatio-temporal information to produce descriptive captions.

There are great potentials in the real-world applications of specificity-controlled video captioning (SCVC). In a human-robot interaction scenario, with the help of SCVC, robots can present descriptions of its perceptual data with a controlled level of detail to humans. For example, a robot docent can produce explanations based on museum visitors' comfort levels [12]. SCVC can also be used to provide video classification or narration to visually impaired users, and the users can control the amount of information they receive.

Figure 1: A comparison between SCVC (ours) and Seq2Seq.
Ground Truth Caption : a man is playing a guitar
Seq2Seq Video Caption : a man is playing a guitar
SCVC at Specificity 1 : a man playing guitar
SCVC at Specificity 2 : a man is playing a guitar
SCVC at Specificity 3 : a man seated is playing the guitar and singing a song

Previous video captioning models lack of the ability to control the amount of detail contained in a caption. They either generate descriptions at an inconsistent level of specificity [20], or produce dense captions describing all major events occurring in a video [10][24], which may not be desirable if the user only needs a brief one-sentence summary of the video.

This paper introduces a specificity-controlled video captioning model that generates a one-sentence description of an open-domain video at a level of granularity specified by the user. We consider video captioning as a translation task from a source video to a target caption, analogous to machine translation (MT), and build SCVC on an encoder-decoder framework [6] specialized in MT tasks. The SCVC model uses a convolutional neural network (CNN) VGG 16 [19] to first extract visual features from video frames then a sequence to sequence model with attention (Seq2Seq-Attn) [1] to generate captions. Feature-wise linear modulation (FiLM) [15] layers are inserted in the decoder of Seq2Seq-Attn to condition the generated caption on a desired specificity level. We draw insights from [25] and use some statistics of a text corpus to compute the specificity level of a caption. Thus the training set consists of tuples of video, caption and caption specificity.

We quantitatively and qualitatively evaluate the efficacy of the SCVC model and show that the produced captions have the desired levels of specificity while maintain the caption quality of the baseline video captioning model that does not support specificity control.

The contributions of this paper are:

1. A domain agnostic Specificity-Controlled Video Captioning (SCVC) model that generates one-sentence descriptions of videos at a level of granularity specified by the user.

2. Two novel evaluation metrics designed for SCVC tasks that measure the caption and specificity fidelity of the generated captions.

2. Related Work

In this section, we summarize prior arts in video captioning, language specificity, neural network conditioning techniques, show some limitations of related work and highlight the novelty of our work.

2.1. Video Captioning

Different from an image, a video consists of a sequence of temporally dependent frames. Thus an effective video captioning model should capture both the spatial structure contained in a single frame and the temporal relations of scenes among frames.

[20] trains an end-to-end model that first extracts feature vectors from video frames using a convolutional neural network (CNN), then feeds the feature vectors into a multilayer LSTM model to generate a caption.

To better capture the temporal dependencies among video frames, [3] adds a boundary-detecting gate in GRU that outputs only hidden states summarizing video substreams between detected boundaries. [10] introduces a model consisting of a CNN for event proposal and an LSTM for event description generation to capture and describe major events occurring in a video.

To improve the quality of single-sentence captions, [23] captures the local and global temporal structures of a video by using 3D convectional neural networks and Bahdanau attention in a CNN-RNN model.

These models do not explicitly control the amount of detail contained in the generated captions. Thus the produced captions may have inconsistent levels of specificity or more details than what a user needs.

In the limited cooking domain, [16] can describe videos in both single sentences and dense captions at the expense of model complexity. Since this approach relies on using hand-centric features for object recognition, it would be hard to generalize it to other domains, where videos do not always contain both human hands. To the best of our knowledge, our work is the first to use a single end-to-end model for open-domain video captioning with explicit specificity control.

2.2. Language Specificity

Language specificity defines the amount of information contained in a sentence.

In a conversational setting, [25] acquires the specificity of a sentence by computing the normalized maximum of inverse word frequencies from the training set and utilizes it as a specificity measure to guide the response generation in the decoder of a Seq2Seq model.

In addition to using some statistics of the text corpus, sentence specificity can also be captured from a text corpus by using a regression model as in [7].

We build upon [25] and define a metric that utilizes sentence length and normalized maximum inverse word frequencies to compute the specificity value of a caption, where word frequencies are acquired from a corpus much larger than the training set to avoid dataset bias.

In a video captioning task, the higher the specificity value is, the more detailed information a caption contains. For example, "A girl is running on a treadmill" is a more specific description than "A person is exercising"

2.3. Specificity Control by Conditional Neural Networks

[15] proposes Feature-wise Linear Modulation (FiLM) to condition arbitrary layers of a neural network by applying an affine transformation to its feature maps, where the scaling and translation parameters are generated from some auxiliary input (e.g. the query question in a visual question answering setting). Although FiLM is developed to improve the performance of structured, multi-step visual reasoning tasks, it has been adapted and proved successful in many other computer vision and natural language processing models [2][8].

To control a certain property of the generated text in neural machine translation, simple concatenation-based approaches work well. [18] controls the politeness of the translation by appending a binary token at the end of the source sentence. [9] appends a token at the beginning of the source sentence to select the target language of the translation from a set of candidate languages.

We examine some simple concatenation-based approaches for our specificity-controlled video captioning task, but they do not work well. We suspect video captioning datasets are multi-modal (i.e. visual and textual), and it is difficult to influence the model output by appending a low dimensional specificity token to a base feature vector of hundreds or thousands dimensions.

3. Method

In this section, we explain in details the metric used to compute specificity level of a caption, the SCVC model and the loss function used for training.

3.1. Specificity Metric

To compute the specificity level of a caption, we first compute a specificity value, represented by a real number ranging from 0 to 1 (inclusive), then discretize the specificity values of captions per video.

The specificity value of a caption c in a text corpus $\mathcal{C}$ is defined as the product of the length and the normalized maximum inverse word frequency (NMIWF) of the caption. We use NMIWF proposed in [25] but multiple it with the sentence length. Sentence length is significant in the specificity computation because longer sentences normally contain more information. If a high specificity value is targeted, the proposed specificity metric encourages SCVC to generate long captions that contains more less frequent words.

$$Specificity(c) = len(c) \times NMIWF(c), \text{ for } c \in \mathcal{C} \quad (1)$$

$$MIWF(c) = \max\{\frac{1}{freq(w)} | \ \forall w \in c\} \quad (2)$$

$$NMIWF(c) = \frac{MIWF(c) - \min\{MIWF(c') | \forall c' \in \mathcal{C}\}}{\max\{MIWF(c') | \forall c' \in \mathcal{C}\} - \min\{MIWF(c') | \forall c' \in \mathcal{C}\}} \quad (3)$$

Instead of computing specificity value on the training set as in [25], we infer the word frequencies from the Google Books Ngram Corpus [11], which contains 155 billion words, to mitigate the dataset bias problem that occurs on a much smaller corpus.

To discretize specificity values, we rank captions per video by their specificity values and equally divide captions into three levels, low, medium and high. The higher the specificity value is, the more details a caption contains. We decide to rank captions and discretize their specificity values per video instead of over the entire caption corpus because some specific words are used more frequently in videos of certain domains. For example, the words cook, stir and fry are normally used in cooking videos but not in sports videos. In Section 3.2, we show that the proposed metric captures various levels of specificity present in the dataset.

3.2. Specificity Metric Validation

We show that the variation of caption specificity present in MSVD can be captured by the proposed specificity metric, and the metric matches closely to human intuition of language specificity.

We randomly select 20 videos from the validation set. Then for each video, we randomly sample a caption of each specificity level measured by the proposed specificity metric. We manually label each of the 3 sampled caption as low, medium or high. Finally, we compare the results of manual labelled specificity level with that computed by the specificity metric and find that $49/60 \approx 81.7\%$ captions receive the same label. All the inconsistently labeled captions are from 5 out of 20 videos.

3.3. Specificity-Controlled Video Captioning (SCVC)

We build our Specificity-Controlled Video Captioning (SCVC) Model (Figure 2) on a Attention-based Sequence to Sequence (Seq2Seq-Attn) model. In addition to the input video, the SCVC model takes in a target specificity level dictating the amount of information the output caption needs to contain.

We randomly sample 80 frames from the input video then feed them in the order of occurrence into a VGG 16, that is pretrained on the ImageNet. The output of the logit layer (4096 dimensional vector) is used as the visual feature representation for each frame. Then this 2D visual embedding of the given video (80×4096) is fed into the encoder of the FiLMed Seq2Seq-Attn to produce a compact representation used as the input to the decoder.

Seq2Seq-Attn is a bidirectional encoder-decoder network consists of gated recurrent units (GRU) [5] and Bahdanau attention mechanism [1]. The size of the hidden units is 512.

FiLM works by first generating the scaling and translation parameters of a set of affine transformations then applying them to the input of each decoder cell. The second input to SCVC, the target specificity level, is used as the auxiliary input to the FiLM generator that produces the parameters. The decoder generates a one-sentence caption describing the input video with the amount of detail specified by the target specificity level.

3.4. Loss Function

During training, we use a separately trained classifier, a feed-forward network with a GRU and a linear layer of size 512, to predict the specificity level of the generated caption. We then use the deviation of the predicted specificity level from the target specificity level as a loss signal, named specificity loss, to enforce the FiLM generator to generate affine transformation parameters that help produce captions close to the target specificity level. The specificity classifier achieves 81% accuracy on the validation set after training 5000 epochs on the training set.

The loss function consists of two parts, caption loss and specificity loss, which are both cross entropy loss between

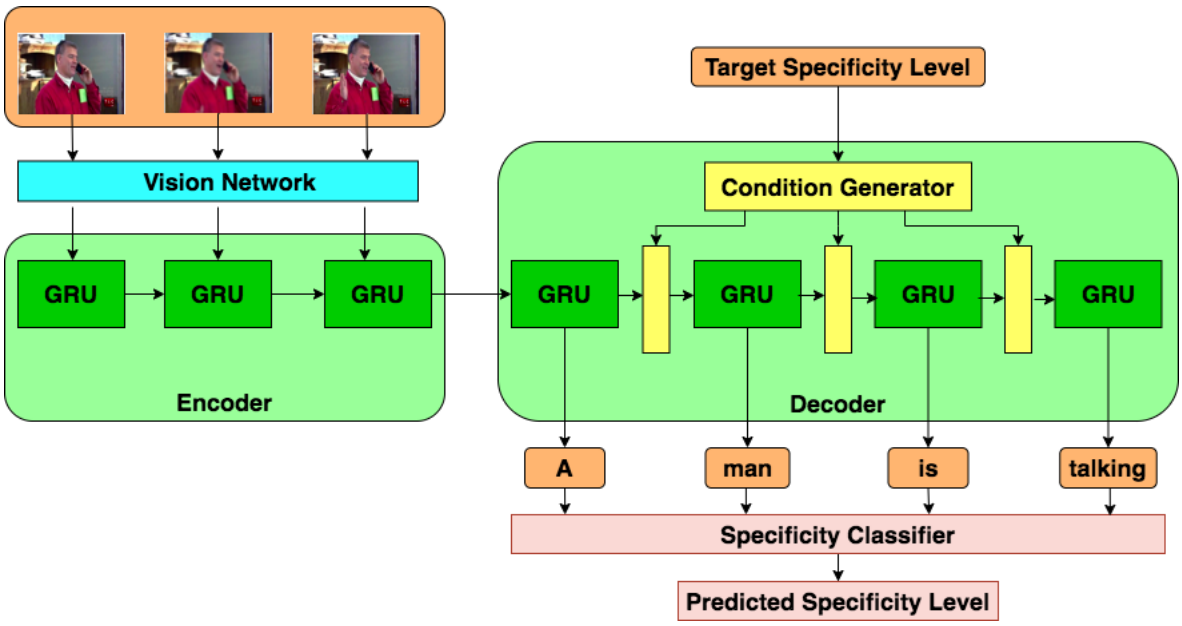

Figure 2: The Specificity-Controlled Video Captioning Model consists of a pretrained vision network (in blue) to extract visual features from video frames, an encoder-decoder network (in green) for caption generation and a conditional mechanism (in yellow) to control the specificity level of produced captions. Then input and output of the model are in orange. The specificity classifier and its output that are used for training are in pink.

the ground truth and predicted values.

$$L(c, \hat{c}, s, \hat{s}) = L_{\text{cap}}(c, \hat{c}) + \alpha L_{\text{spec}}(s, \hat{s}) \tag{4}$$

α is a hyperparameter which we set to be 10 since we observe that at the beginning of the training process, the L_{cap} is about 10 times larger than L_{spec}, and we would like the loss function to penalize SCVC for generating captions at incorrect specificity levels as much as inaccurate captions.

4. Microsoft Video Captioning Dataset (MSVD)

We use a standard video captioning dataset, Microsoft Video Captioning Dataset (MSVD) [4]. Each video is annotated with a set of one-sentence captions of varying specificity levels. In the following, we provide some statistics of this dataset and prove its validity for the specificity-controlled video captioning task.

4.1. Statistics

MSVD contains 1970 videos ranging from 41 to 1799 frames. The videos are divided into training, validation and test sets with 1500, 100 and 370 videos and their corresponding captions, respectively.

The number of captions per video ranges from 18 to 81 with a median of 40 captions. The caption length varies from 1 to 141 words with a median of 8. The vocabulary contains 13,394 unique words.

For each caption, we compute a specificity level. Together, MSVD contains 80,839 video-caption-specificity triplets, and 62,060 (76.7%) of them are unique.

4.2. Dataset Validation

To perform specificity-controlled video captioning, the dataset is required to contain captions of various levels of specificity.

We show that MSVD has this property by plotting the histogram of the log specificity values of captions in the text corpus. We observe that the histogram has the shape a normal distribution with mean specificity value at approximately 0.00673794699.

We plot the word frequency histograms of the 200 to 3000 most frequent words in sentences at each specificity level (bottom row of Figure 4. We see distinct patterns of words usage by sentences that are at low, medium and high specificity level.

5. Experiments and Results

In this section, we first validate baseline VC which we build SCVC on, then show experiment results that demonstrate the efficacy of the SCVC model and compare it with 3 ablation models and an baseline video captioning model.

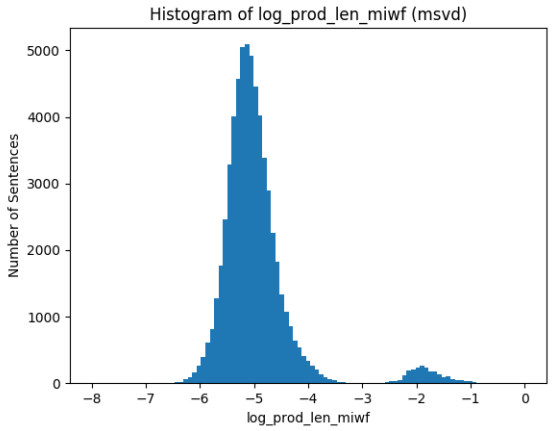

Figure 3: The histogram of the log specificity values of captions in the text corpus

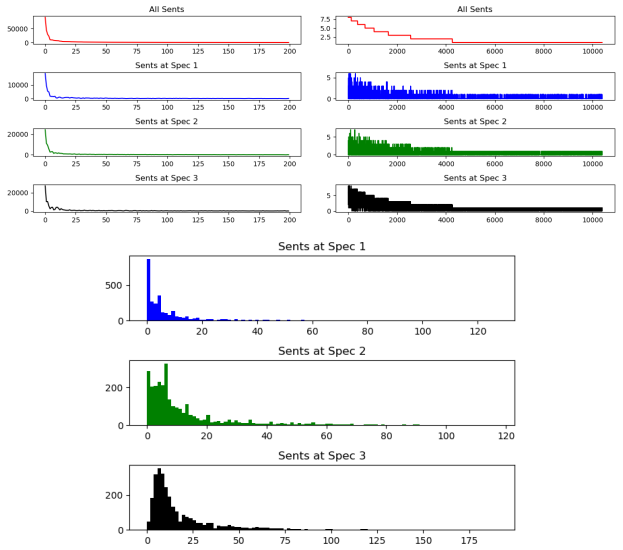

Figure 4: Plots from top to bottom are word frequencies of all sentences and sentences at low, medium and high specificity levels. Figure in the top row are word frequencies of top 200 most frequent words and words after the 3000th most frequent word. Figure in bottom row are the words frequencies of 200-3000 most frequent words.

5.1. Baseline Video Captioning Model

Our SCVC model is built on top of a baseline Seq2Seq-Attn video captioning model. We prove that the baseline model produces captions conditioned on the input videos instead of memorizing word usage of the caption corpus. We feed random frames sampled from a normal distribution into Seq2Seq-Attn and observe that the output captions are random and have low evaluation scores (BLEU_4=0.1677, METEOR=0.0839, CIDEr=0.0622, ROUGE_L=0.4531).

5.2. Quantitative Evaluation of Caption and Specificity Fidelity

We train the SCVC model using the training set of MSVD for 6000 epochs and validate it every 100 epochs. The total, specificity and caption loss curves for the training and validation set are in Figure 5.

Because the task of specificity-controlled video captioning (SCVC) is different from conventional video captioning (VC), the metrics used to evaluate VC are not suitable for SCVC. VC evaluation metrics (e.g. BLEU, METEOR, CIDEr, ROUGE_L) measure caption correctness but not specificity correctness of the generated captions. Thus we design two metrics specifically for SCVC and compare the performance of SCVC with that of a baseline model using SCVC and VC metrics.

5.2.1 Specificity-Controlled Video Captioning Evaluation Metrics

We use mean average precision (mAP) to define two SCVC evaluation metrics to measure the semantics and specificity fidelity of the generated captions.

For every video in the validation set, we randomly sample an equal number of captions from the corpus as the number of the ground truth captions. Thus for each video, we have a set of good (ground truth) and bad (randomly sampled) captions. We then use teacher forcing to produce video captions and compute their perplexities of a trained model. For an effective SCVC model, the two metrics would rank the captions generated from the good caption set higher than that from the bad caption set in terms of semantics and specificity correctness.

We compute an average precision (AP) score over the set of generated captions for individual video, then average the AP scores of all videos to get a mAP measure of the model. The scikit-learn AP function has two inputs, a list of binary labels and a list of scores [14]. We use perplexities of the generated captions as the scores and compute binary labels differently to measure caption and specificity correctness.

To measure semantic correctness, we define the binary label of a caption to be 1 if it is generated from the ground truth caption and 0 otherwise. For specificity correctness, binary label represents whether the specificity level of generated captions equals the target specificity level.

We compare the performance of the SCVC model with 3 ablation models, each of which is trained with an identical target specificity level, 1, 2 and 3 respectively. The ablation models represent an elimination of specificity control used in SCVC. The results are show in Table 1. As shown, the SCVC model trained with ground truth specificity level,

computed by the metric introduced in Section 3.1, outperforms the ablation models while maintaining the same caption fidelity.

5.2.2 Video Captioning Evaluation Metrics

Since the goal of SCVC is to generate captions of various specificity levels while matching the caption quality of the baseline video captioning model, we show that after ablating the specificity input, SCVC generates descriptions as accurate as the Seq2Seq-Attn video captioning model.

The results of caption quality by two models are comparable as shown in Table 2.

5.3. Qualitative Evaluation

For each validation video, we ask the trained SCVC model to generate three captions, one at each specificity level, low, medium and high.

We inspect the generated captions and find a consistent trend that the lengths of captions increase as we increase the target specificity level, and more specific words usually occur in sentences with higher specific levels. A comparison of captions generated by specificity-controlled video caption and baseline Seq2Seq-Attn video caption models are shown in Table **??** .

We observe 8 out of 100 examples in the validation set as the target specificity level increases, the captions generated by SCVC become wrong, which might indicate the SCVC sometimes focus too much on getting the right specificity than correct caption in high specificity. In one case, as target specificity level increases, SCVC adds more details, which might not be correct, to the correct low specificity level captions. In some other cases, as target specificity levels increase, SCVC decides uses rare words and generates a longer sentence that is wrong or partially correct.

In contrast, in some cases, as target specificity levels increase, parts of the sentences generated by SCVC start to become right, which means that specificity control could sometimes help improve the quality of produced caption

When the three captions generated by SCVC for a video are completely wrong, they are often still at different levels of granularity, which suggests the conditional network tries to generate more complicated captions as we increase the target specificity level.

5.4. Computing Resources

We used 2 Nvidia GeForce 2080Ti to train and validate all the aforementioned models.

6. Conclusion and Future Work

This paper describes a Specificity-Controlled Video Captioning (SCVC) model that generates captions of an open-domain video at a level of granularity specified by the user. SCVC influences the output of a Seq2Seq-Atten language model by adding conditional layers in its decoder.

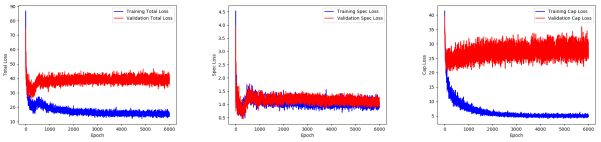

Figure 5: The total, specificity and caption loss curves (from left to right) of true-spec. The blue curves represent training loss and the red curves are for validation loss.

One limitation of current approach is that as the target specificity level increases, SCVC would try to add in more details whihc might not be correct to generated captions. A significant performance boost might be achieved by condition the visual model (e.g. 3D ResNet) on the target specificity input, so it produces visual attentions focusing on regions of video frames that is most useful for specificity-conditioned language generation.

Next we would like to train SCVC end-to-end and evaluate its performance on larger datasets, like MSR-VTT-2017 [22][13], VATEX [21] and movie descriptions [17].

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli, and Edward Grefenstette. Learning to follow language instructions with adversarial reward induction. *arXiv preprint arXiv:1806.01946*, 2018.

[3] Silvia Cascianelli, Gabriele Costante, Thomas Alessandro Ciarfuglia, Paolo Valigi, and Mario Luca Fravolini. Full-gru natural language video description for service robotics applications. *IEEE Robotics and Automation Letters*, 3:841–848, 2018.

[4] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[5] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] Yifan Gao, Yang Zhong, Daniel Preotiuc-Pietro, and Junyi Jessy Li. Predicting and analyzing language specificity in social media posts. 2019.

	SCVC	SCVC-1	SCVC-2	SCVC-3	VC	VC-1	VC-2	VC-3
Semantic Correctness	0.690	0.777	0.769	0.618	0.756	0.813	0.800	0.668
Specificity Correctness	0.963	0.940	0.530	0.460	0.275	0.120	0.110	0.770
Sem AND Spec Correctness	0.728	0.732	0.425	0.292	0.193	0.095	0.086	0.513

Table 1: From left to right are the SCVC model trained using ground-truth captions and their corresponding specificity levels as targets evaluated on the entire validation set (SCVC), the subset containing only captions at specificity 1 (SCVC-1), 2 (SCVC-2), 3 (SCVC-3) and the baseline Seq2Seq-Attn VC model trained using only ground-truth captions evaluated on on the entire validation set (VC), the subset containing only captions at specificity 1 (VC-1), 2 (VC-2), 3 (VC-3)

	BLEU@4	METEOR	CIDEr	ROUGE_L
SCVC	0.3976	0.1393	0.4913	0.5507
VC	0.3582	0.1339	0.5609	0.5419
GitHub	0.38			

Table 2: Caption quality of SCVC and Seq2Seq-Attn video captioning models, and BLEU score reported on the GitHub repository of the baseline Seq2Seq-Attn model.

[8] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017.

[9] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.

[11] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics, 2012.

[12] Ruikun Luo, Sabrina Benge, Natalie Vasher, Grace VanderVliet, John Turner, Maani Ghaffari, and X Jessie Yang. Toward an interactive robot docent: Estimating museum visitors' comfort level with art.

[13] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195. Springer, 2014.

[17] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017.

[18] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, 2016.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[21] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591, 2019.

[22] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.

[24] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on*

	Total	Spec = 1	Spec = 2	Spec =3
SCVC Cap Correct	64/150=0.427	30/50=0.6	16/50=0.36	18/50=0.32
SCVC Cap Correct	89/150=0.593	33/50=0.66	27/50=0.54	30/50=0.6
VC Cap Correct	38/50=0.74			

Table 3: More Tolerate Human evaluation of captions generated by SCVC and VC in terms of caption and specificity correctness.

	Total	Spec = 1	Spec = 2	Spec =3
SCVC Cap Correct	64/50=0.427	30/50=0.6	16/50=0.36	18/50=0.32
SCVC Spec Correct	141/150=0.94	47/50=0.94	44/50=0.88	50/50=1.0
VC Cap Correct	38/50=0.74			

Table 4: Human evaluation of captions generated by SCVC and VC in terms of caption and specificity correctness.

computer vision and pattern recognition, pages 4584–4593, 2016.

[25] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117, 2018.