Language-Driven Region Pointer Advancement for Controllable Image Captioning

Annika Lindh and **Robert Ross** and **John D. Kelleher**
ADAPT Research Centre
School of Computer Science,
Technological University Dublin,
Kevin Street, Dublin 8, IRELAND
{annika.lindh;robert.ross;john.d.kelleher}@tudublin.ie

Abstract

Controllable Image Captioning is a recent sub-field in the multi-modal task of Image Captioning wherein constraints are placed on which regions in an image should be described in the generated natural language caption. This puts a stronger focus on producing more detailed descriptions, and opens the door for more end-user control over results. A vital component of the Controllable Image Captioning architecture is the mechanism that decides the timing of attending to each region through the advancement of a region pointer. In this paper, we propose a novel method for predicting the timing of region pointer advancement by treating the advancement step as a natural part of the language structure via a NEXT-token, motivated by a strong correlation to the sentence structure in the training data. We find that our timing agrees with the ground-truth timing in the Flickr30k Entities test data with a precision of 86.55% and a recall of 97.92%. Our model implementing this technique improves the state-of-the-art on standard captioning metrics while additionally demonstrating a considerably larger effective vocabulary size.

1 Introduction

Image Captioning brings together the two fields of Computer Vision and Natural Language Generation into a task where the model needs to translate an input image into an appropriate natural language text description. The task leaves some ambiguity regarding which parts of the image should be mentioned and which ones can be excluded. This has led to a common problem where models tend to generate overly generic descriptions that seem to focus more on the category of the image than on its individual content (Devlin et al., 2015; Madhyastha et al., 2018).

Recently, Cornia et al. (2019) introduced Controllable Image Captioning as a new sub-task of Image Captioning with a stronger focus on image details. In this task, the input to the model is an image along with bounding box coordinates for a sequence of regions (where each region can consist of one or more bounding boxes) that must be explicitly described in the candidate caption. Thus, in contrast to standard Image Captioning, a generic candidate caption would not meet the criteria of a suitable caption even if it contained no factual errors; this is reflected in the evaluation process where the candidate caption is only compared to those ground-captions that share the same sequence of regions. Fig. 1 shows an image and two corresponding captions from the Flickr30k dataset (Young et al., 2014) along with complementary data from Flickr30k Entities (Plummer et al., 2017) which provides annotations that link entities in the captions to region bounding boxes in the corresponding image.

In terms of practical use, a Controllable Image Captioning model provides more flexibility and user-control over the generated captions, without having to retrain the model. Since the region selection process is not entangled with the caption generation components, the former can be swapped out to adapt to different scenarios (e.g. when applied to social media images, it could target regions where a facial recognition system has tagged friends of the user). Furthermore, it opens up the possibility of adhering to individual user preferences regarding the amount and type of details to describe – this is a feature that could provide real benefits to blind and low-vision users (Stangl et al., 2020).

Proceedings of the 28th International Conference on Computational Linguistics, pages 1922–1935
Barcelona, Spain (Online), December 8-13, 2020

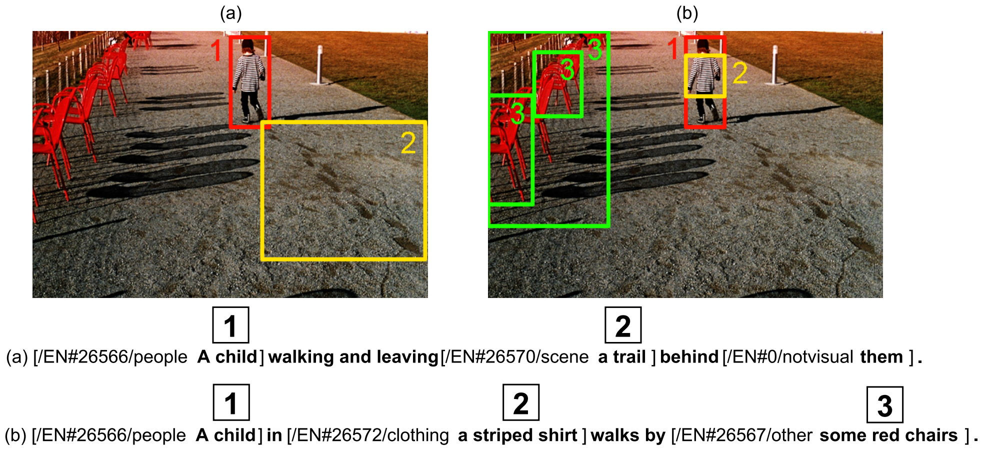

(a)
(b)
1
2
1
2
3
3
3
1
2
3
(a) [/EN#26566/people A child] walking and leaving [/EN#26570/scene a trail] behind [/EN#0/notvisual them] .
(b) [/EN#26566/people A child] in [/EN#26572/clothing a striped shirt] walks by [/EN#26567/other some red chairs] .

Figure 1: Image and captions from Flickr30k (Young et al., 2014) with regions and text annotations from Flickr30k Entities (Plummer et al., 2017). Square brackets indicate entity annotations. Some entity regions consist of multiple bounding boxes.

When it comes to model architecture, a key component in Controllable Image Captioning is the mechanism that advances the region pointer at the appropriate time so that each requested region is sufficiently described. In previous work, this has been approached as its own prediction task in parallel to the prediction of the next word (Cornia et al., 2019). However, in this paper, we demonstrate that the timing of the region pointer advancement is strongly linked to the sentence structure in the ground-truth and thus could likely be predicted as if it were a natural language token. Hence, we propose a novel approach to the region pointer advancement mechanism that directly leverages the language model to generate a unique NEXT-token as part of the caption generation process. We implement this in a Controllable Image Captioning model where we demonstrate its effectiveness by measuring how often the NEXT-token is predicted in agreement with the ground-truth. Furthermore, we measure the model's overall performance on standard Image Captioning metrics where it outperforms the current state-of-the art on the Flickr30k (Young et al., 2014) benchmark test. Additionally, we observe that our model demonstrates a considerably larger effective vocabulary size than the current state of the art model, thus enabling our model to describe a greater variety of visual content.

We make our code and trained model publicly available for future work to build on.[1]

2 Related Work

Controllable Image Captioning builds directly on the standard Image Captioning task and thus involves a similar set of components, with the main difference being the role of visual attention. In the early neural models for Image Captioning, visual information was typically extracted into a fixed-size vector representation of the full image, which would be provided as input either at the start or at each recurrent timestep in the generation process (e.g. Kulkarni et al. (2013), Fang et al. (2015), Vinyals et al. (2015)). These early neural models were found to produce largely generic captions with output similar to retrieval-based models (Devlin et al., 2015; Madhyastha et al., 2018).

In recent years, dynamically generated attention has become the standard method to mitigate generic captions through enhanced visual grounding by providing the model with only the visual information relevant to the current timestep; multiple variations of dynamic attention have been explored, from the soft and hard attention proposed by Xu et al. (2015) to the region proposal network used by Anderson et al. (2018).

In Controllable Image Captioning, the locations of the regions of interest are included in the input to the model (as shown in Fig. 1). To extract meaningful features from these regions, previous work in Controllable Image Captioning has successfully used the region features from Anderson et al. (2018) that

[1] https://github.com/AnnikaLindh/Controllable_Region_Pointer_Advancement

were developed for the standard captioning task but whose object-focused training method corresponds well to the object-focused region annotations in the Flickr30k Entities dataset (Plummer et al., 2017).

Since the *selection and ordering* of the regions is fixed, the challenge instead lies in predicting the appropriate *timing* of attending to each region. Since the current task requires that the output incorporates exactly the requested regions, the model must ensure that each part of the generated caption is sufficiently grounded in its corresponding visual region and that the generated sequence of words is not terminated before the complete region sequence has been described.

To predict the timing of attending to each region, Cornia et al. (2019) implement a region pointer along with a mechanism to predict, at each timestep, whether this pointer should be incremented or not. The prediction follows a multi-step process that includes an additional LSTM layer to model the *attention state* of the current caption chunk, extended with an additional output layer to model the *end* of this chunk, as well as a chunk-shifting gate module that outputs the probability of incrementing the region pointer, based on: the chunk's attention state, the chunk's end state and the visual features for each of the requested regions. A potential weakness in this implementation is the structure of connection between the word prediction LSTM, the attention LSTM and the sampling of the next word: the chunk-shifting gate (which comprises an earlier step in this module) must make its decision *before* the next word has been sampled. Thus, it must predict whether the next word *is going to be* the last word of the chunk rather than predicting whether the current word *is* the end of the chunk. This complicates the task of the chunk-shifting gate which must rely on incomplete information about the chunk when predicting its end-point. Furthermore, it would be difficult for the chunk-shifting gate to anticipate the result of any non-deterministic method for sampling the next word.

In the following section, we argue that a more elegant solution for advancing the region pointer is not only possible but also preferable.

3 Region Pointer Advancement

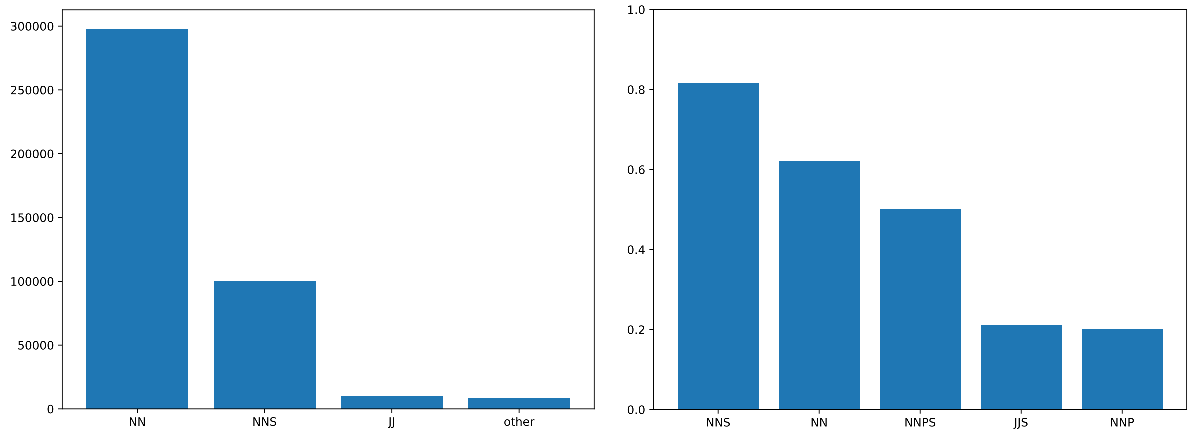
Region pointer advancement training relies on separating full captions into chunks that each relate to a visual region. We use the same chunking method as Cornia et al. (2019) on the Flickr30k Entities (Plummer et al., 2017) captions, meaning that each chunk starts with the word immediately following the end of the previous chunk, and ends with the last word in the following entity annotation that has at least one bounding box associated with it; entity annotations without bounding boxes are not taken into account.

Fig. 1 shows two example captions with entity annotations enclosed in brackets, where each entity annotation includes an entity ID, an entity type and the noun phrase associated with that entity. After chunking, example *a* consists of three chunks: 1) *a child*, 2) *walking and leaving a trail* and 3) *behind them*. The visual region associated with each chunk refers to the average-pooled features of all bounding boxes associated with its entity. Since the third chunk does not have an associated entity, we associate it with the empty (zero-vector) region during training.

Since the end of a chunk is related to the position of its entity's noun phrase, it seems reasonable to suspect that the end of a chunk is correlated to parts-of-speech (PoS). We explore two potential correlations after employing automatic PoS tagging on the training data, using the Python implementation of the Brill tagger (Brill, 1992) found in the TextBlob[2] library. Fig. 2a shows the frequency of each PoS tag assigned to the last word of a chunk; unsurprisingly, the last word of a chunk is most commonly tagged as a noun. Further, Fig. 2b shows the positive predictive value (i.e. precision, which takes into account how often each PoS tag appears in the training data) for indicating the end of a chunk among the top five most predictive PoS tags. Again, nouns are most commonly associated with ending a chunk, though their positive predictive value varies from 81.6% for plural nouns (NNS) to 20.0% for singular proper nouns (NNP).

From the statistics shown in Fig. 2, it is clear that the end of a chunk (and thus, the region pointer advancement timing) is correlated to the sentence structure. Thus, we propose to inject a special language token (NEXT) to mark the end-of-chunk events, treating it in a similar way to the beginning-of-sequence

[2] https://textblob.readthedocs.io



(a)

(b)

Figure 2: PoS tag statistics for the final word of chunks in the training data. (a) shows the total counts of each tag appearing at the end of a chunk, while (b) shows the positive predictive value for indicating the end of a chunk of the top five most predictive tags.

(BOS) and end-of-sequence (EOS) events. The two examples in Fig. 1 thus become: (a) *BOS a child NEXT walking and leaving a trail NEXT behind them EOS*, and (b) *BOS a child NEXT in a striped shirt NEXT walks by some red chairs NEXT EOS*. With the added NEXT-tokens, our model no longer requires additional steps or learnable layers to predict the region advancement timing since this is treated as a natural language property and predicted by the language model (as any other word). Unlike the method used by Cornia et al. (2019) (described in section 2), our prediction mechanism has access to the most recently generated word regardless of sampling method. Additionally, during the immediately following timestep, the NEXT-token becomes the previous word, and thus our language model is explicitly informed that a new chunk should begin.

To evaluate our proposed region pointer advancement method, we implement it as part of a Controllable Image Captioning model (described in section 4) and test it on the region sequence scenario (Cornia et al., 2019) where each example's candidate caption is compared only to those ground-truth captions that share the same region sequence (i.e. the same regions in the same order).

4 Model Architecture

Our architecture, shown in Fig. 3, is similar to a typical Image Captioning model based on the encoder-decoder design with visual attention. The recurrent unit is a two-layer Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) module with a hidden size of 1024 units (using the implementation details of Zaremba et al. (2015)). The input i_t at each timestep t is the weighted concatenation of the previous word embedding w_{t-1} and the current visual region's embedding v_t, with the weight balance α_t between them being generated according to equations 1-2:

$$\alpha_t = \sigma(h_{t-1} \cdot W_\alpha + b_\alpha) \tag{1}$$
$$i_t = [\alpha_t \cdot w_{t-1}, (1 - \alpha_t) \cdot v_t] \tag{2}$$

where W_α, b_α are the learnable weights and bias of a single linear layer. The hidden state from the LSTM is passed through a single learnable linear layer to produce the logits over the full vocabulary.

Our word embeddings are randomly initialized learned vector representations of size 1024, while the visual embedding is a vector of size 2053. The first 2048 features are extracted from the final pooling layer (pool5) of a Bottom-Up network (Anderson et al., 2018) where the region's bounding box(es) are used as the region proposals; in the case where a region is made up of multiple bounding boxes, we average-pool their features into a single vector representation of size 2048. The last 5 features consist

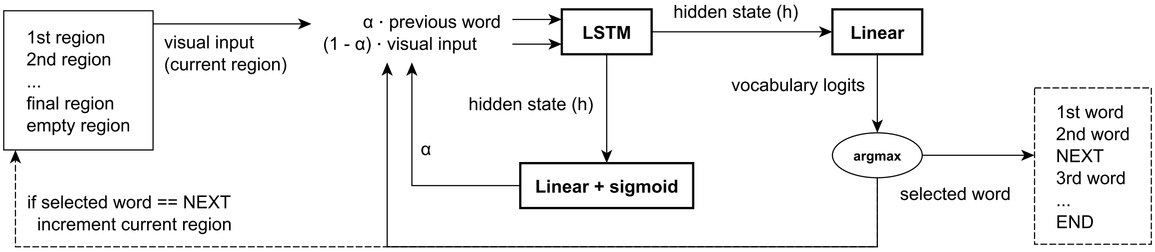


Figure 3: Model architecture and flowchart for recurrent steps. Solid line arrows indicate direct information flow, while the dashed line arrow indicates an action taken based on the information.

of: the count of bounding boxes that made up this region, as well as the bounding boxes' maximum and minimum x and y coordinates relative to the total image size.

A full-image sized bounding box is used to extract the overall visual features of the image; these features are passed through two separate learnable linear layers to initialize the LSTM's hidden and cell states respectively. (These initialization steps are not featured in Fig. 3.)

After extracting the visual features from the Bottom-Up network and initializing the LSTM's hidden state and cell state (as described above), the model recursively generates a caption word-by-word until the EOS-token is produced. The process for a single step in this recursion is as follows: the LSTM's previous hidden state is passed through a single linear layer with a sigmoid activation unit to produce the α_t weight according to equation 1; this weight (and its complement) decides how much attention the model should pay to the text input and visual input respectively. The text input at each timestep is the learned word embedding for the previously generated token (or the BOS-token for the first step), while the visual input consists of the features representing the region currently indicated by the region pointer. The region pointer initially points to the first region in the requested region sequence, and is subsequently incremented each time the NEXT-token is produced by the model; if the NEXT-token is generated when no further requested regions remain, the region pointer instead points to a zero-vector region (i.e. the empty region). Once the text and visual parts of the input have been concatenated according to equation 2, the resulting full input is passed through the LSTM module, followed by a single linear layer to produce the logits over the full training vocabulary from which the next word is sampled. This process continues until the EOS-token is produced which marks the end of the caption.

Due to the discrete-valued region pointer, the model will only receive visual input from a single region at any one timestep. However, since the LSTM's memory states are not reset between chunks, it is possible for the model to retain information about previous regions along with information about the word embeddings of previously generated words when transitioning from one chunk to the next within a single caption.

During training, the previous token refers to the ground truth token of the previous step rather than the previously generated token. The special tokens NEXT and EOS are trained in the same way as the word tokens in the vocabulary; however, the region pointer advancement follows the ground-truth NEXT-tokens only, and generating EOS early will not prevent the model from training on the remaining part of the caption.

5 Experiments

In our experiments, we use the Karpathy splits (Karpathy and Fei-Fei, 2017) of the Flickr30k (Young et al., 2014) dataset, giving us 29,000 images for training, 1014 for validation and 1000 for the test set; each image is associated with 5 human-annotated captions. In Controllable Image Captioning, a unique example is defined by an image along with a unique sequence of regions; thus, each image in the Flickr30k dataset corresponds to between 1 and 5 unique examples in our experiments (at most 1 per caption if they all use different region sequences). Thus, while the number of unique images remain the same, the number of examples in our splits become: 120,667 examples for training, 4208 for validation and 4148 for testing.

We follow the standard procedure for pre-processing captions: all captions were lower-cased and stripped of punctuation, and we replace rare words (less than 5 occurrences in the training set) by an UNK token in the training data; no words were replaced in the test set. The captions were then split into chunks based on the human-annotated entity annotations (as described in section 3), where each chunk ends with a visually grounded entity's associated noun phrase (or the last word of the caption respectively). During test time, each example consists of an image, a (possibly empty) sequence of regions and one or more ground-truth captions to measure success against. The Bottom-Up network (Anderson et al., 2018) used for the visual embeddings was trained on a custom split of the Visual Genome (Krishna et al., 2016) dataset (after standard pre-training on ImageNet (Russakovsky et al., 2015)) to avoid an overlap between the images in the Bottom-Up net's training set and the images from our model's test set. The weights of the Bottom-Up net were frozen and remained fixed during training while the word embedding features were learned end-to-end along with the rest of the model.

We implement our model (described in section 4) using the PyTorch[3] framework. We use a batch size of 100, a learning rate of $1e^{-5}$ and a dropout of 0.7 for both the previous word embedding and the LSTM's layer connections.[4] To prevent overfitting, the model was evaluated on the validation set every 10 epochs, and the checkpoint with the best CIDEr (Vedantam et al., 2015) metric was selected. During inference on the test set for the final results, the models were prevented from generating the UNK token by setting its probability to zero. If the end-of-sequence (EOS) token was generated before all regions had been used, it was instead interpreted as a NEXT-token. If there were no more regions when a NEXT-token was generated, the visual input was referred to the empty (zero-vector) region.

All learned parameters were trained using the cross-entropy loss over the generated word sequence, to minimize the negative log probability of the ground-truth words from each caption, per equation 3:

$$loss = \frac{1}{T}\sum_{t=1}^{T} - \log P(w_t|w_{t-1}, v_t, S_t) \tag{3}$$

where w_t and w_{t-1} are the current and the previous ground-truth words, v_t is the visual input and S_t is the current memory state of the model at time t. All learnable weights were initialized to random uniform floats in range [-0.1, 0.1] except biases which are initialized to 0.0. The word embedding features were given a random uniform initialization in range [-1.0, 1.0].

5.1 Ablation Tests

The purpose of our region pointer advancement method is to enable our model to generate strongly grounded language by attending the appropriate visual region at each timestep. Meanwhile, this also allows our language model to learn valuable information about the chunk structure of the captions, which may, in itself, be a useful tool for improved language generation. To better assess the effect of appropriately timed region attention, we train an ablation version of our model (called *Ours, average-pooled* in the tables), which learns the same chunk structure in the text but does not have access to the individual region features. Instead, we replace the individual region embeddings with the average-pooled features[5] from the current example's full region sequence. Thus, the average-pooled model receives visual information for all regions in the relevant sequence, but is unable to attend to individual regions at each timestep. As with the full model, the average-pooled model is informed whether it has described the full number of regions, and it likewise receives the empty region as input when the full number of region chunks have been generated.

6 Results and Analysis

To test our full model's ability to learn the appropriate timing of the region pointer advancement using our proposed method, we measure how often the NEXT-token is predicted in agreement with the ground-

[3] https://pytorch.org/

[4] The LSTM dropout uses the method from Zaremba et al. (2015) which does not apply dropout to the recurrent connections.

[5] The last five features (describing each region's bounding boxes) are replaced by zeros.

Table 1: Standard metrics: B=BLEU, R-L=ROUGE-L, C=CIDEr, M=METEOR, S=SPICE. *a* indicates mean results across 6 runs; *b* indicates single-run model results from Cornia et al. (2019). Models above the horizontal middle line report results without finetuning.

Model	B-1	B-2	B-3	B-4	R-L	M	C	S
Ours, average-pooled	37.49	23.18	14.98	9.96	33.38	15.12	64.51	18.70
Ours, full model[a]	**41.77**	**27.14**	**18.42**	**12.83**	**38.97**	**17.33**	**87.25**	22.17
95% Confidence Interval	±0.15	±0.15	±0.11	±0.07	±0.12	±0.08	±0.52	±0.14
SCT (CE)[b]	33.62	22.47	15.68	11.25	36.86	15.42	74.52	23.45
SCT (CIDEr)[b]	39.26	25.79	17.58	12.36	38.84	16.58	83.72	23.45
SCT (CIDEr NW)[b]	40.44	26.51	17.97	12.52	38.93	16.75	83.99	**23.50**

truth from the test set when guiding the model with the correct previous word from the ground-truth data at each timestep (i.e. in the teacher-guided scenario). We find that our model generates the NEXT-token in agreement with the ground-truth to a precision of 86.55% and a recall of 97.92%, thus establishing that our region pointer advancement method works well in practice and as such is likely to contribute to the successful training of our model. In the rest of our experiments, the ground-truth captions are not known during inference, and the previous word refers to the model's previously sampled word.

To evaluate how our model performs on the Controllable Image Captioning task, we measure the model's performance on the standard captioning metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016), using the implementation from the speaksee[6] tool. The first four of these are *n*-gram based metrics, with CIDEr being developed specifically for Image Captioning, while the other three have been borrowed from the machine translation and summarization fields. Among this class of metrics, CIDEr and METEOR have been associated with the highest correlation to human evaluation scores on the standard Image Captioning task (Bernardi et al., 2016); they provide a measure of both the fluency and content agreement of the generated captions in relation to the ground truth. Like CIDEr, SPICE was also developed as an Image Captioning metric, but instead measures the overlap between inferred scene graph tuples from the candidate and ground-truth captions, and is thus intended to measure the accuracy of the objects and their relations in the candidate captions.

We compare our results to the Show Control and Tell (SCT) model from Cornia et al. (2019) who introduced the Controllable Image Captioning task. We compare to three versions of the SCT model: SCT (CE), SCT (CIDEr) and SCT (CIDEr NW). SCT (CE) is trained with the standard cross-entropy loss, while the other two are first trained on the cross-entropy loss and then further finetuned using Reinforcement Learning. SCT (CIDEr) is finetuned towards the CIDEr metric, while SCT (CIDEr NW) is finetuned on a combination of CIDEr and a region alignment score (defined by equation 13 from Cornia et al. (2019)) based on word embedding similarity between nouns of corresponding caption chunks. All three SCT models use beam sampling with a beam size of 5. We present the results from our model using argmax sampling and standard cross-entropy loss training and leave further finetuning up to individual applications. For our full model we report the mean across 6 runs along with 95% Confidence Intervals.

Table 1 shows that our full model outperforms the SCT models on all standard metrics except SPICE (with only the ROUGE-L score of SCT (CIDER NW) falling within our 95% Confidence Interval), suggesting that our generated captions are more similar to the ground-truth captions. Interestingly, our model outperforms all SCT models on the CIDEr score, including the two SCT models that have been finetuned specifically towards this metric. All models except our average-pooled version produce shorter captions than the ground-truth's average 12.4 words, with our full model being the second closest with an average of 12.1 words. The SCT (CE) model has the shortest average caption length of 9.4 words, while the finetuned SCT (CIDER NW) increases this to 11.4.

While the standard captioning metrics provide us with a measure of the fluency and content overlap

[6] https://github.com/aimagelab/speaksee

Table 2: Diversity statistics: Diversity = distinct captions, Novelty = captions not seen in the training set, Vocab = total number of unique words in the generated captions, Length = average number of words in the captions. *a* indicates mean results across 6 runs.

Model	Diversity %	Novelty %	Vocab	Length
Ground Truth	99.96	99.70	4247	12.4
Ours, average-pooled	90.96	96.02	1042	12.4
Ours, full model[a]	**95.36**	96.83	**1200**	12.1
95% Confidence Interval	±0.32	±0.13	±36	±0.1
SCT (CE) (Cornia et al., 2019)	86.24	95.98	935	9.4
SCT (CIDEr) (Cornia et al., 2019)	89.98	97.79	538	10.9
SCT (CIDEr NW) (Cornia et al., 2019)	92.06	**98.07**	502	11.4

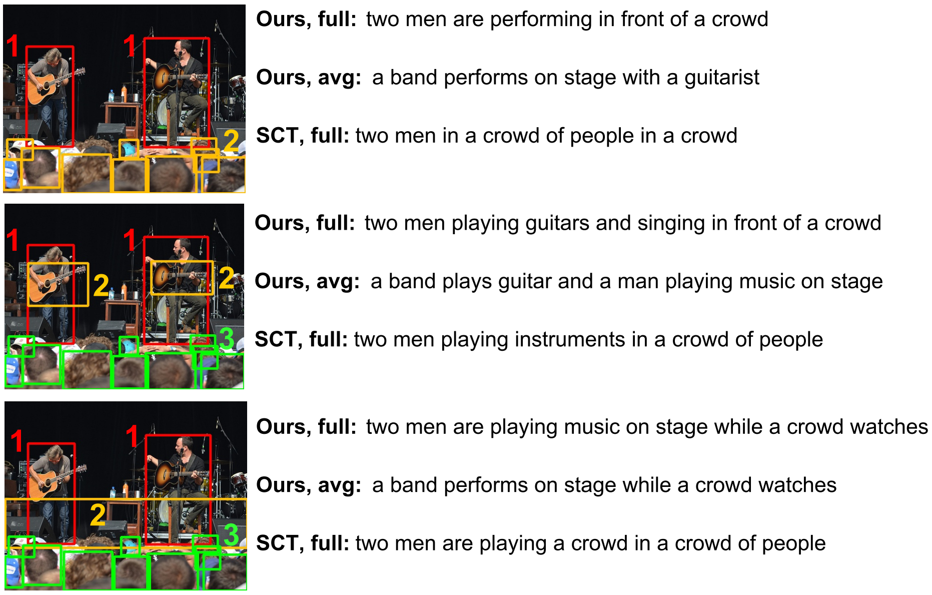
with the ground-truth, they do not directly penalize a model's tendency towards repetitive captions (Lindh et al., 2018). We expect a model with a well-functioning region pointer advancement timing to encourage more diverse and detailed captions due to better alignment with the visual input – in comparison, if the pointer advancement lags behind the language-driven transition to the next chunk, then it would be necessary for the language model to start generating the next chunk of words without access to the relevant visual input, possibly by relying more on word co-occurrence rates in the text. To test this, we employ three metrics proposed by Lindh et al. (2018): *diversity* (proportion of distinct candidate captions), *novelty* (proportion of candidate captions not found in the training set) and *effective vocabulary size* (total number of unique words across all candidate captions).

From Table 2 we can see that all models, including our average-pooled baseline, perform well on both the Diversity and Novelty metrics, confirming that the Controllable Image Captioning setting promotes captions that are not generic or repetitive. Our full model generates the highest number of distinct captions (95.36%), while the fully finetuned SCT (CIDEr NW) model generates the highest number of captions that were not seen in the training set (98.07%). When it comes to generating captions with a varied vocabulary our full model has by far the largest effective vocabulary size at 1200 unique words – ahead of SCT (CE) at 935 unique words and more than double that of the finetuned SCT models at 538 and 502 unique words respectively. The considerable decrease in SCT's vocabulary size after finetuning (despite an increase in their average caption lengths) might indicate an unwanted side-effect of CIDEr optimization, possibly encouraging a preference for common *n*-grams while not sufficiently rewarding uncommon words.

Finally, from our ablation test we can tell that the average-pooled features combined with knowledge about the appropriate number of chunks is sufficient to produce acceptable results, despite using the same visual features at each timestep. However, while the average-pooled model performs acceptably, our full model is still clearly ahead on the standard captioning metrics. Thus, the results indicate that our full model does indeed learn a region pointer advancement timing that is useful for learning to generate visually grounded language.

A possible explanation for the relatively good results of the average-pooled ablation model could be that the average-pooled model learns to internally keep track of the current chunk number along with memorizing a typical sentence structure (e.g. by learning to describe people in the first chunk, followed by their attributes, and describing their activity in a later chunk). Another possibility is that it learns a type of content planning similar to older encoder-decoder models without attention, with the NEXT-token aiding in learning the sentence structure and with the additional benefit of having the empty region input indicating when to end the sequence.

Fig 4 shows the generated captions for three different region sequences on the same image. (More examples can be found in the appendix.) Overall, we found that all models were capable of producing reasonable and detailed descriptions, but that the SCT models seemed more likely to produce common-



Ours, full: two men are performing in front of a crowd

Ours, avg: a band performs on stage with a guitarist

SCT, full: two men in a crowd of people in a crowd

Ours, full: two men playing guitars and singing in front of a crowd

Ours, avg: a band plays guitar and a man playing music on stage

SCT, full: two men playing instruments in a crowd of people

Ours, full: two men are playing music on stage while a crowd watches

Ours, avg: a band performs on stage while a crowd watches

SCT, full: two men are playing a crowd in a crowd of people

Figure 4: Variations in captions on different region sequences from the same image. *SCT, full* is the fully finetuned SCT (CIDEr NW) model from Cornia et al. (2019).

sense relationship errors (e.g. *playing a crowd* from the bottom example in Fig 4). This difference can likely be explained as an effect of our model's region pointer advancement timing being strictly tied to the sentence structure, thus allowing it to attend to the next region before generating the words that tie the two regions together. In contrast, if the model was to delay its region pointer advancement with as little as a single timestep, it would need to start generating relationship words based solely on the first of the two regions with no knowledge of the second.

7 Conclusion and Future Work

Based on the strong correlation between sentence structure and region-related chunks in the training data's captions, we proposed a language-driven method of region pointer advancement in Controllable Image Captioning. We have implemented our proposed method in a Controllable Image Captioning model where it demonstrates a precision of 86.55% and a recall of 97.92%. Our full model outperforms the current state of the art model on the standard metrics, including CIDEr, despite using only the cross-entropy loss whereas the current state-of-the-art relies on finetuning towards the CIDEr metric.

Additionally, we find that our model has an effective vocabulary size that is more than double that of the current state-of-the-art, suggesting that our model is more capable of learning and generating uncommon words.

We have demonstrated that our method for region pointer advancement works well in the vision-to-text context. However, its implementation could be applied to any sequence-to-sequence tasks where structural chunks in the input data (e.g. image regions) can be related to structural chunks in the output (e.g. natural language sentence chunks); some possible applications would be Speech-to-Text, Machine Translation or the standard Image Captioning task when combined with a region selection and sorting mechanism.

For the task of Controllable Image Captioning, we would encourage future work to consider complementary metrics such as caption diversity and effective vocabulary size (alongside the standard captioning metrics) to better understand a model's capacity to generate unique descriptions for each unique input. Additional metrics to specifically measure the adherence to an ordered region sequence would be welcome.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 17/RC-PHD/3488 at the ADAPT SFI Research Centre at Technological University Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *arXiv:1607.08822 [cs]*, July. arXiv: 1607.08822.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research*, 55(1):409–442, January.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, ANLC '92, pages 152–155, Trento, Italy, March. Association for Computational Linguistics.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8299–8308.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language Models for Image Captioning: The Quirks and What Works. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2:100–105.

H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. 2015. From captions to visual concepts and back. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, June.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November.

A. Karpathy and L. Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv:1602.07332 [cs]*, February.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903, December.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 605–612, Stroudsburg, PA, USA. Association for Computational Linguistics.

Annika Lindh, Robert J. Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D. Kelleher. 2018. Generating Diverse and Meaningful Captions. In Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, Lecture Notes in Computer Science, pages 176–187. Springer International Publishing.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2018. The role of image representations in vision to language tasks. *Natural Language Engineering*, 24(3):415–439, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318, Philadelphia, July. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, 123(1):74–93, May.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. ”Person, Shoes, Tree. Is the Person Naked?” What People with Vision Impairments Want in Image Descriptions. In *CHI 2020*, pages 1–13. ACM, April.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, June.

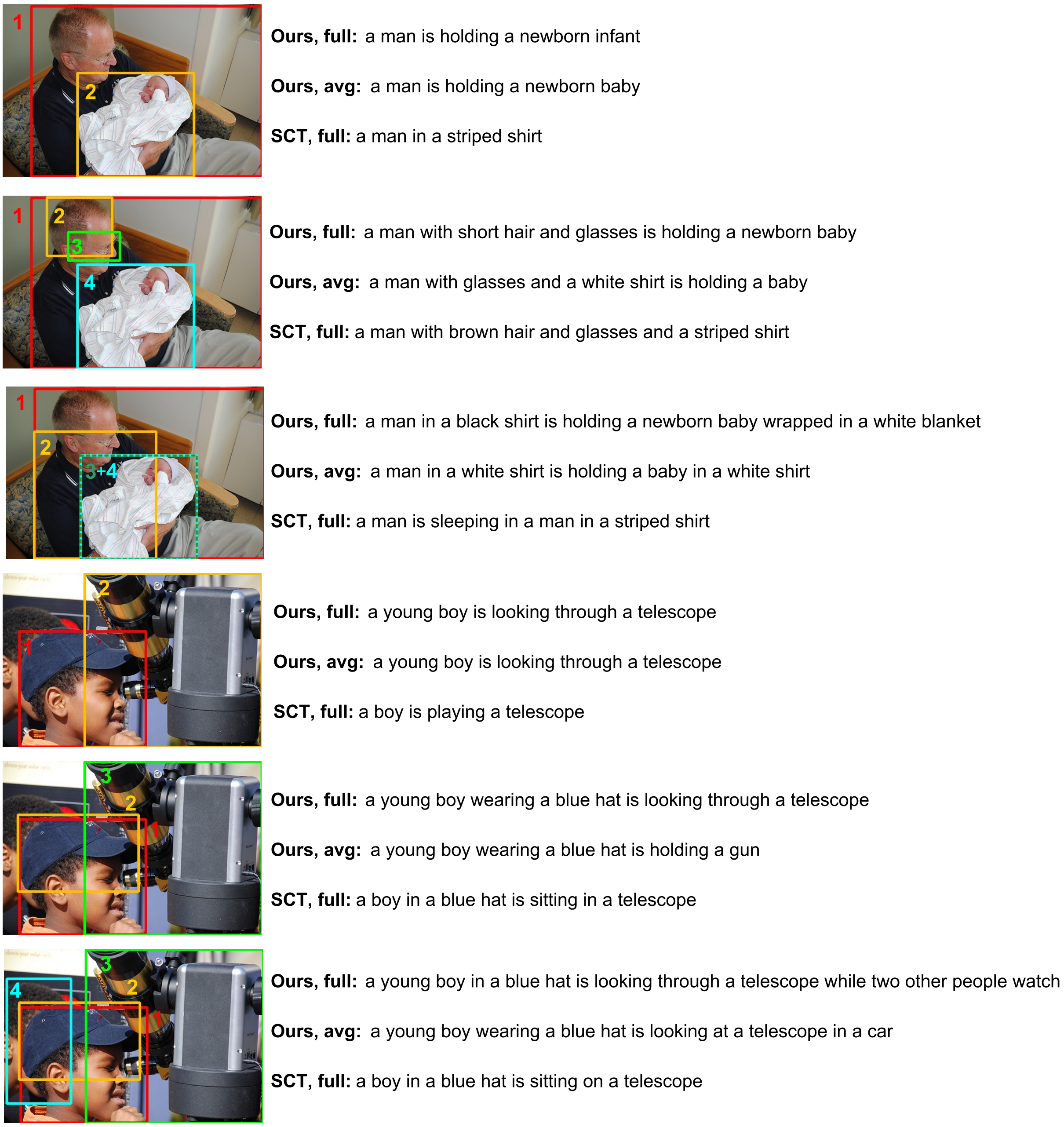
Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, June. ISSN: 1063-6919.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, July. PMLR.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2(0):67–78, February.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2015. Recurrent Neural Network Regularization. *arXiv:1409.2329 [cs]*, February.

Appendix. Additional Caption Examples.



Ours, full: a man is holding a newborn infant

Ours, avg: a man is holding a newborn baby

SCT, full: a man in a striped shirt

Ours, full: a man with short hair and glasses is holding a newborn baby

Ours, avg: a man with glasses and a white shirt is holding a baby

SCT, full: a man with brown hair and glasses and a striped shirt

Ours, full: a man in a black shirt is holding a newborn baby wrapped in a white blanket

Ours, avg: a man in a white shirt is holding a baby in a white shirt

SCT, full: a man is sleeping in a man in a striped shirt

Ours, full: a young boy is looking through a telescope

Ours, avg: a young boy is looking through a telescope

SCT, full: a boy is playing a telescope

Ours, full: a young boy wearing a blue hat is looking through a telescope

Ours, avg: a young boy wearing a blue hat is holding a gun

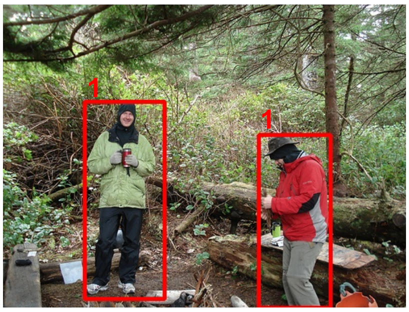
SCT, full: a boy in a blue hat is sitting in a telescope

Ours, full: a young boy in a blue hat is looking through a telescope while two other people watch

Ours, avg: a young boy wearing a blue hat is looking at a telescope in a car

SCT, full: a boy in a blue hat is sitting on a telescope

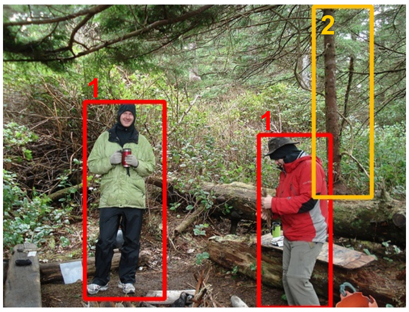
Figure 5: Additional examples of generated captions. *SCT, full* is the fully finetuned SCT (CIDEr NW) model referenced in Section 6.



Ours, full: two people are hiking in a forest

Ours, avg: a man is walking through a forest

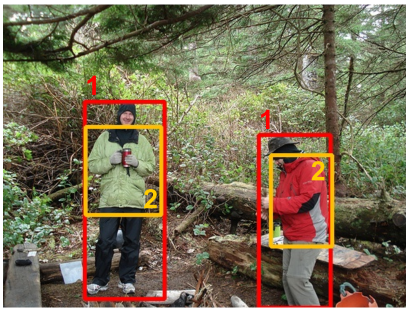
SCT, full: two people are standing in a man



Ours, full: two people are walking down a tree lined path

Ours, avg: a man is standing in a tree

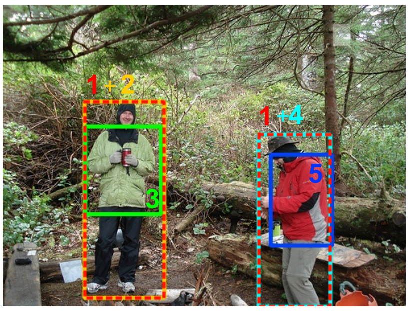
SCT: two people are standing in a tree with a tree



Ours, full: two people in jackets are hiking in the woods

Ours, avg: a man in a red jacket is walking in the wilderness

SCT, full: two people in a red jacket and green jackets

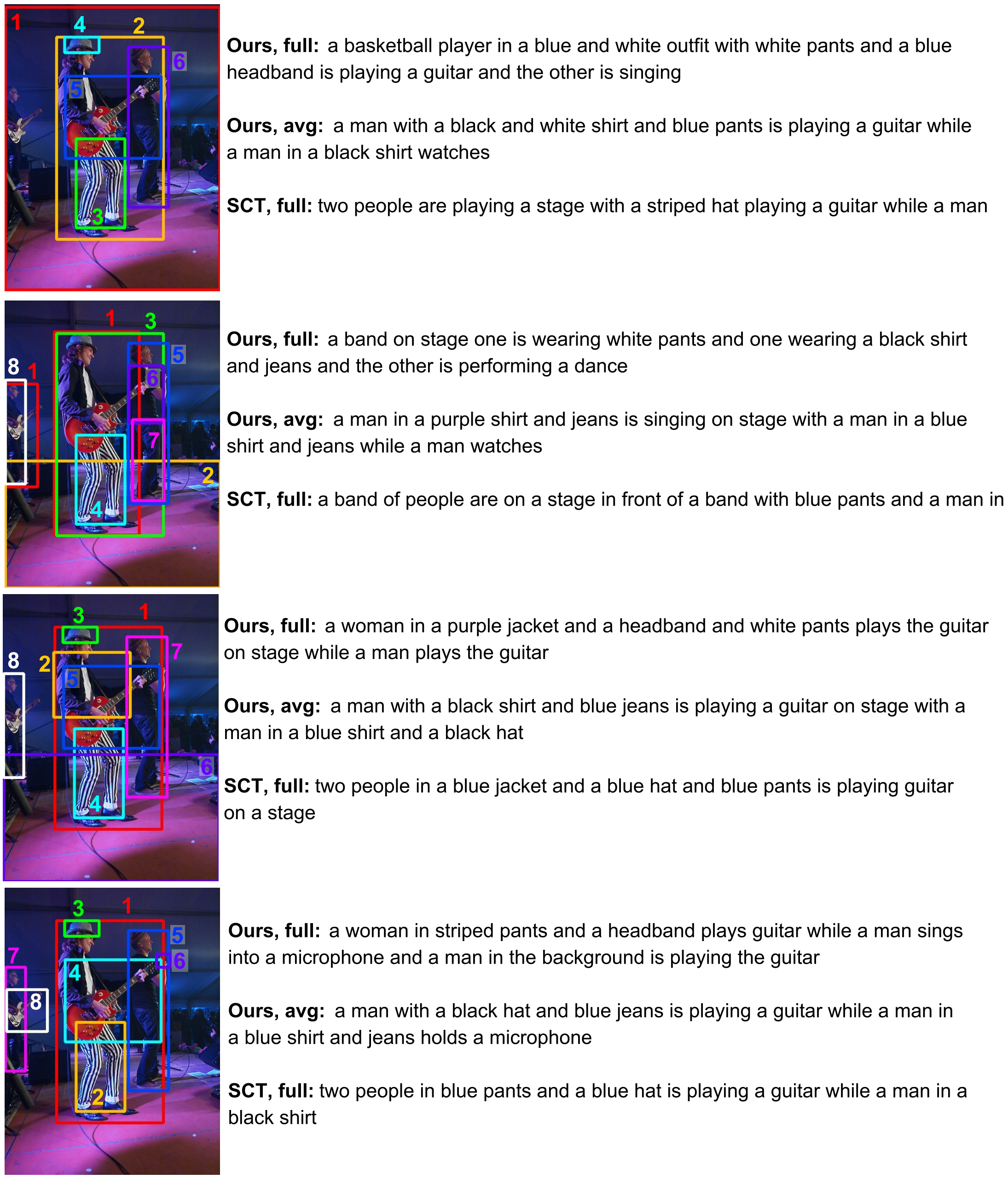


Ours, full: two people one in a green jacket and the other in a red shirt are walking in the grass

Ours, avg: a man in a red jacket is looking at a woman in a green jacket and a boy in a field

SCT, full: two people are standing in a man in a green jacket and a man in a red jacket

Figure 6: Additional examples of generated captions. *SCT, full* is the fully finetuned SCT (CIDEr NW) model referenced in Section 6.



Ours, full: a basketball player in a blue and white outfit with white pants and a blue headband is playing a guitar and the other is singing

Ours, avg: a man with a black and white shirt and blue pants is playing a guitar while a man in a black shirt watches

SCT, full: two people are playing a stage with a striped hat playing a guitar while a man

Ours, full: a band on stage one is wearing white pants and one wearing a black shirt and jeans and the other is performing a dance

Ours, avg: a man in a purple shirt and jeans is singing on stage with a man in a blue shirt and jeans while a man watches

SCT, full: a band of people are on a stage in front of a band with blue pants and a man in

Ours, full: a woman in a purple jacket and a headband and white pants plays the guitar on stage while a man plays the guitar

Ours, avg: a man with a black shirt and blue jeans is playing a guitar on stage with a man in a blue shirt and a black hat

SCT, full: two people in a blue jacket and a blue hat and blue pants is playing guitar on a stage

Ours, full: a woman in striped pants and a headband plays guitar while a man sings into a microphone and a man in the background is playing the guitar

Ours, avg: a man with a black hat and blue jeans is playing a guitar while a man in a blue shirt and jeans holds a microphone

SCT, full: two people in blue pants and a blue hat is playing a guitar while a man in a black shirt

Figure 7: Additional examples of generated captions. *SCT, full* is the fully finetuned SCT (CIDEr NW) model referenced in Section 6.