2023

Show, Prefer and Tell: Incorporating User Preferences into Image Captioning

Annika Lindh
Technological University Dublin, annika.lindh@tudublin.ie

Robert J. Ross
Technological University Dublin, robert.ross@tudublin.ie

John Kelleher
Technological University Dublin, john.kelleher@tudublin.ie

Show, Prefer and Tell: Incorporating User Preferences into Image Captioning

Annika Lindh, Robert Ross, John D. Kelleher
ADAPT Centre, School of Computer Science, Technological University Dublin
D07 EWV4
Ireland
{annika.lindh; robert.ross; john.d.kelleher}@tudublin.ie

ABSTRACT

Current image captioning models produce fluent captions, but they rely on a one-size-fits-all approach that does not take into account the preferences of individual end-users. We present a method to generate descriptions with an adjustable amount of content that can be set at inference-time, thus providing a step toward a more user-centered approach to image captioning.*

CCS CONCEPTS

• **Accessibility technologies** • **Natural language generation** • **User centered design** • **Scene understanding**

KEYWORDS

Image Captioning, Assistive Technology, Deep Learning

ACM Reference format:

Annika Lindh, Robert Ross, John D. Kelleher. 2023. In *Proceedings of ACM SAC Conference, Tallinn, Estonia, March 27-31, 2023 (SAC'23)*, 4 pages. DOI: 10.1145/3555776.3577794

1 INTRODUCTION

Image Captioning (IC) is the task of generating natural language descriptions for images. Models encode the image using a convolutional neural network (CNN) and generate the caption via a recurrent model or a multi-modal transformer. Success is measured by the similarity between generated captions and human-written "ground-truth" captions, using the CIDEr [14], SPICE [1] and METEOR [2] metrics. While incremental gains have been made on these metrics, there is a lack of focus on end-user opinions on the amount of content in captions. Studies with blind and low-vision participants have found that lack of detail is a problem [6, 13, 17], and that the preferred amount of content varies between individuals [13], as do individual opinions on the trade-off between correctness and adding additional content with lower confidence [9]. We propose a more user-centered approach with an adjustable amount of content based on the number of regions to describe. We demonstrate that our model can generate fluent captions across a range of settings. The generated captions along with the source code and visual examples are made publicly available online[1].

2 RELATED WORK

Most similar to our work is Deng et al. [5] who approximately control the number of words within a set of binned ranges. A limitation is that their model must be re-trained toward each length setting. Another limitation is that by controlling number of words, a longer caption could describe more objects, or use more words to describe the same objects, or introduce unnecessary repetitions.

Controllable Image Captioning (CIC) instead enables control through a set of regions [4, 8] or an inferred scene-graph [18]. This does not rely on pre-defined settings and is thus more flexible. Current CIC models typically rely on ground-truth regions during inference which limits their practical usability. An exception is Zhong et al. [18] who experiment with auto-selection from a scene-graph; though this was limited to shorter descriptions of single sub-graphs. Chen et al. [3] use automatically detected regions, but rely on verbs from the ground-truth captions during inference.

We propose a model that incorporates preferences of amount of content, without relying on ground-truth during inference.

3 MODEL ARCHITECTURE

The model is composed of five components executed in sequence:

1. **Region Detection** extracts bounding boxes and features.
2. **Region Grouping** combines boxes, e.g. grouping many person-boxes into a larger region representing a crowd.
3. **Region Selection** selects the highest quality regions.
4. **Region Ordering** plans the region order in the caption.
5. **Caption Generation** generates caption while ensuring that each selected region is included in the planned order.

Each component was trained independently and are described in this section, with additional details on our GitHub page.

SAC '23, March 27-March 31, 2023, Tallinn, Estonia

ACM ISBN 978-1-4503-9517-5/23/03.
https://doi.org/10.1145/3555776.3577794

[1] https://github.com/AnnikaLindh/show-prefer-tell

The **Region Detection** component produces up to 300 bounding boxes through the Region Proposal Network (RPN) component of a Faster R-CNN [12], along with corresponding visual features. We use the official Caffe[2] weights from Lindh et al. [8] and load them into the PyTorch[3] framework. Before passing the image to the network, the mean pixel values are subtracted from the current image (rather than using the average pixel means from all training data which is the default setting); manual inspection showed that this improved detections of people with darker skin tones. Any boxes identified as the *background* class are discarded. Remaining boxes are assigned one of the categories listed in Table 1, by summing the top three category probabilities and selecting the highest. These categories come from Flickr30k Entities [11] but the RPN uses more fine-grained classes, so conversion was made by recursively matching the WordNet[4] Synsets listed in Table 1. If no match was found, the category called *other* was assigned. Regions with a summed score higher than 0.3 are accepted as candidates, and Non-Maximum Suppression (NMS) is applied to discard possible duplicate regions of the same category, using a threshold Intersection of Union (IoU) of 0.3 between bounding box pairs.

The **Region Grouping** step decides whether regions from the same category should be grouped. It uses a network of two fully connected Rectified Linear Unit (ReLU) layers followed by a single-unit sigmoid output layer, where a value above 0.5 triggers two boxes to be grouped. Regions that are adjacent in either horizontal or vertical spatial order can be grouped, which can lead to a chain of multiple regions being grouped. The input is their relative size, x and y center distances, and a one-hot encoding of the category. The features of grouped regions are averaged, while the bounding box is set to fully encompass all the boxes. The network was trained on all pairs of adjacent ground-truth boxes with the same category. True pairs were given a label of 1.0; and negative pairs were labeled as 0.0. Negative pairs were given a weight of 0.99 during training to account for class imbalance.

Region Selection gives a score of 0.0 to 1.0 to each region and selects the top regions up to the number of requested regions. Image features are passed through a fully connected layer with 20 ReLU output units; this output is concatenated to another 11 features: the normalized x and y distances between the region and image centers, area size relative to the image, a one-hot encoding of the category, and the category's probability score. These features are passed through 2 Leaky ReLU layers and a final single-unit sigmoid layer. During training, a label of 1.0 was given to RPN detections with a minimum IoU of 0.4 to ground-truth; else a label of 0.0. Since some detections could be grouped later, a simulated grouping step was added where detections with an IoU of 0.3 or above were combined and compared to ground-truth regions again; if this IoU was 0.4 or higher, then those detections were given a label of 1.0.

For **Region Ordering**, two methods were compared. The first is Sinkhorn ordering from Cornia et al. [4] which takes the visual features of each region, the normalized x and y coordinates, and the GLoVe [10] embedding of the predicted class; we used the categories from Table 1 instead RPN classes. A pre-defined maximum number of regions is required, which was set to 10. The second method is a rule-based ordering as follows:

1. Add the largest region in the first category with at least one region, using the category priority: *people*, *animals*, *instruments*, *vehicles*, *other, clothing/bodyparts*.
2. If that region was *person* or *animal*, add all overlapping *clothing* regions from largest to smallest, followed by all *bodyparts* regions from largest to smallest.
3. Regardless of region type from step 1, add all remaining overlapping regions from largest to smallest.
4. Use Manhattan distance between region center-points to find the closest region from the one in step 1, then repeat the process as if this region was chosen in step 1.

Table 1: Flickr30k Entities categories with corresponding WordNet Synsets.

Category	WordNet Synset
people	person.n.01
animals	animal.n.01
instruments	musical_instrument.n.01
clothing	clothing.n.01
bodyparts	external_body_part.n.01
vehicles	vehicle.n.01
other	-

For **Caption Generation**, we use the controllable captioner from Lindh et al. [8]. The input at each step is the current region's visual features, normalized coordinates, and an integer to mark how many detections were grouped. When the model generates a NEXT-token, the region pointer advances to the next region in the list [8]. If the model generates the END-token before the final region has been consumed, it is automatically converted into a NEXT-token to prevent the caption from ending too early.

4 EXPERIMENT DESIGN

Image-caption pairs from Flickr30k [16] were used together with Flickr30k Entities [11] which extends the dataset with bounding box coordinates of regions that are mentioned in the captions, with matching entity markers in the text to indicate when they are mentioned. The Karpathy splits [7] were used, providing training, validation and test sets of 29,000, 1014 and 1000 images respectively, each with 5 English human-written captions. Figure 1 shows the frequency of captions by number of regions in the training data, after removing the three examples with zero regions. As can be seen, most training examples have only 2 or 3 regions.

We evaluated our model's ability to produce captions for any number of regions N up to $N = 7$, as beyond this number few ground-truth examples exist. Since previous models relied on some ground-truth input, two ablation models were created to understand the impact of this information:

[2] https://caffe.berkeleyvision.org/
[3] https://pytorch.org/
[4] https://wordnet.princeton.edu/

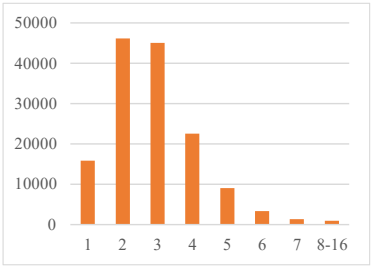

Figure 1: Frequency of captions by number of regions.

- **CIC**: the model from Lindh et al [8] with ground-truth bounding boxes, grouping and ordering.
- **Matched**: uses the same RPN as the full model, but it matches the RPN detections to the ground-truth grouping, selection and ordering. Selection is done through greedy matching with a minimum IoU of 0.3. For grouped ground-truth regions, it first compares to the full region; if this fails and at least two of the region's partial boxes can be matched to RPN detections, then those detections are grouped and selected.

The ablation models use the first *N* regions from examples with a number of regions > *N*. Unique examples were generated to avoid duplicate examples after truncating on the current *N*. Additionally, we evaluate the score of the ground-truth captions, referred to as GT. Each GT caption is only evaluated on the other ground-truth captions, i.e. the same caption is excluded from the references.

Two evaluation modes were used: *All* uses all human-written captions for an image as references, while *Same* uses only those with the same *N*. GT scores were assessed for CIDEr, METEOR and SPICE on the two modes, to understand how stable each metric is at different values of *N*, as shown in Figure 2. *CIDEr All* was very sensitive to *N*, with heavy penalties for including additional content; this was less pronounced for *CIDEr Same* where only ground-truth captions with the same *N* were used. *METEOR All* remains stable at higher *N* values and thus provides a good complement to *CIDEr Same*. SPICE follows METEOR's trend but is slightly less stable. All three metrics struggle on *N*=1 when mode=*All*. Results presented in the next section use the settings *CIDEr Same* and *METEOR All*. When comparing two or more models, evaluation was restricted to those images which all models produced captions for. For models that sometimes produce multiple captions per image due to multiple ground-truth annotations, (i.e. GT, CIC and Matched), the score for each image is the mean score of all caption for that image, thus ensuring that each image receives the same weight. Note that GT evaluation for mode=*Same* can only be carried out on images with at least two human-written captions of the same *N* value.

Table 2: Generated captions for Flickr30k images with IDs 354017707 (top three) and 230486268 (bottom three).

Regions	Captions
1	A man is sitting in a subway.
3	A man in a gray jacket is reading a newspaper.
5	A man in a gray jacket and hat and red scarf is reading a newspaper.
1	A little boy is playing in the sand.
3	A little boy in a red shirt is playing in the sand.
5	A little boy in a red shirt is playing in the sand with a shovel and a brown hat.

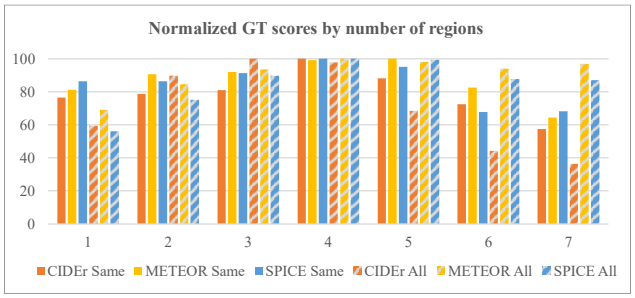

Figure 2: Effect of number of regions on GT scores. Scores are normalized across all region numbers for each metric.

5 RESULTS AND DISCUSSION

To be practically useful, the model must be capable of detecting and selecting enough regions for various image types. Our model was able to generate captions with the full number of requested regions for more than 90% of the images for $N <= 4$ while dropping off toward 50.6% $N = 7$. Grouping results in fewer regions, leading to a coverage of 75.1% at $N = 4$ down to 18.6% at $N = 7$. Caption lengths were in the same range as ground-truth lengths except for $N <= 2$ where the generated captions were 1 word shorter.

Figure 3 *(left)* compares scores with and without grouping, and between rule-based (RB) and Sinkhorn (SH) ordering. Grouping is not preferred with the exception of *CIDEr Same* at the lowest and highest *N* values, though $N = 7$ should be interpreted with some caution as it only had 12 examples. SH ordering seems to be somewhat preferred over RB, with some exceptions at higher *N*, possibly explained by the inability of SH to consider hierarchical relationships between regions. Based on these results, we suggest enabling grouping only at $N = 1$ and to use RB ordering for $N >= 5$. Table 2 shows examples of captions at $N = 1$, 3 and 5 with these settings. For clarity, the remaining score comparisons use SH ordering and no grouping at all *N* values in the Full model.

To put the scores into context and to understand the impact of ground-truth input, evaluation was carried out on examples shared between the Full-NoGroup-SH model and the three ablation models, with results shown in Figure 3 *(right)*. Scores for Full-NoGroup-SH differ from Figure 3 *(left)* since each experiment evaluates only the examples shared between all models. Results for $N > 4$ were left out in this comparison due to few shared examples with GT. As expected, more ground-truth information leads to higher scores, confirming the suitability of the evaluation method. The greatest score difference is seen when switching from ground-truth captions to generated captions, while further restrictions to ground-truth information appears to have smaller, incremental effects. The gap between our full model and Matched is similar to the gap between Matched and CIC. On *CIDEr Same*, our full model performed on par with Matched at $N <= 3$.

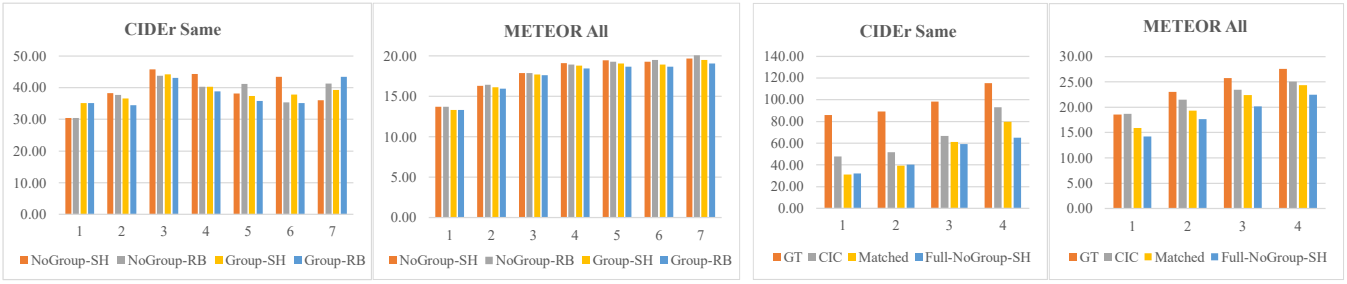

Figure 3: Metrics scores by number of regions, comparing grouping and ordering settings *(left)* and ground-truth *(right)*.

5.1 Error Modes

Manual inspection of the captions and regions revealed two main error modes: 1) *object misclassifications*, e.g. interpreting a shadow as a person, and 2) *unexpected region ordering* causing confusion for the captioner. When the latter caused misplaced modifiers, this was mitigated by RB ordering, except for the cases where *clothing* and *bodyparts* regions were misclassified as *other*.

Finally, for $N = 1$, the selected region would often be a large region containing the main actor, but the captioning would focus on the scenery within that region, leading to awkward captions such as *"A large water is being sprayed."*.

6 CONLUSION AND FUTURE WORK

Automatic descriptions for images can assist blind and low-vision users in consuming information that is otherwise inaccessible. Current solutions often provide insufficient details and do not take into account the varied preferences among individual end-users. To address these issues, we have proposed a model that follows a more user-centered approach by allowing a customizable amount of content in the captions. This bridges the gap between traditional and controllable captioning models, and performs relatively well in comparison to baselines with access to ground-truth information.

Beyond our model results, the comparison of METEOR, CIDEr and SPICE scores for ground-truth captions with different numbers of regions gives insight into the length-sensitivity of these metrics. CIDEr seems to punish captions with more details, which may be a cause for concern when fine-tuning models toward this metric.

ACKNOWLEDGMENTS

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

[1] Anderson, P., Fernando, B., Johnson, M. and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *Computer Vision – ECCV 2016* (Oct. 2016), 382–398.

[2] Banerjee, S. and Lavie, A. 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*. (2005), 65–72.

[3] Chen, L., Jiang, Z. and Liu, W. 2021. *Human-like Controllable Image Captioning with Verb-specific Semantic Roles*.

[4] Cornia, M., Baraldi, L. and Cucchiara, R. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 8299–8308.

[5] Deng, C., Ding, N., Tan, M. and Wu, Q. 2020. Length-Controllable Image Captioning. *Computer Vision – ECCV 2020*. A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. Springer International Publishing. 712–729.

[6] Gleason, C., Pavel, A., McCamey, E., Low, C., Carrington, P., Kitani, K.M. and Bigham, J.P. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA, Apr. 2020), 1–12.

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

[8] Lindh, A., Ross, R.J. and Kelleher, J.D. 2020. Language-Driven Region Pointer Advancement for Controllable Image Captioning. *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona, Spain (Online), Dec. 2020), 1922–1935.

[9] MacLeod, H., Bennett, C.L., Morris, M.R. and Cutrell, E. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA, May 2017), 5988–5999.

[10] Pennington, J., Socher, R. and Manning, C. 2014. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (2014), 1532–1543.

[11] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*. 123, 1 (May 2017), 74–93.

[12] Ren, S., He, K., Girshick, R. and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems 28*. C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, eds. Curran Associates, Inc. 91–99.

[13] Stangl, A., Morris, M.R. and Gurari, D. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Apr. 2020), 1–13.

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

[15] Yan, K., Ji, L., Luo, H., Zhou, M., Duan, N. and Ma, S. 2021. Control Image Captioning Spatially and Temporally. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online, Aug. 2021), 2014–2025.

[16] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. 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 (Feb. 2014), 67–78.

[17] Zhao, Y., Wu, S., Reynolds, L. and Azenkot, S. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments. *Proceedings of the ACM on Human-Computer Interaction*. 1, CSCW (Dec. 2017), 121:1-121:22.

[18] Zhong, Y., Wang, L., Chen, J., Yu, D. and Li, Y. 2020. Comprehensive Image Captioning via Scene Graph Decomposition. *Computer Vision – ECCV 2020*. A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. Springer International Publishing. 211–229.