Towards Real-Time Image Captioning using Crowdsourcing and Computer Vision

Divya Ramesh & Bradford A. Folkens
CloudSight Inc.
5455 Wilshire Blvd., Suite 1111, Los Angeles, CA - 90036

MOTIVATION

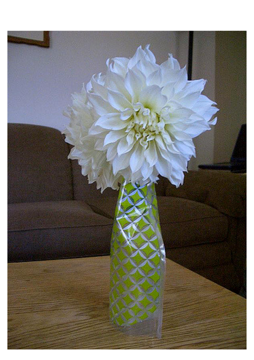
1. How can we improve accessibility for the visually impaired?
2. Can visual captioning be made real-time yet reliable using a combination of both machine and human intelligence?

CHALLENGES

- How do we validate a machine generated caption in real-time?
- Existing measures such as BLEU, METEOR, ROUGE and CIDEr are sensitive to n-gram overlap.
- They do not take into account the visual attributes present in the image.
- SPICE compares visual attributes but is sensitive to scene graphs.
- All methods require at least one reference caption to determine validity.

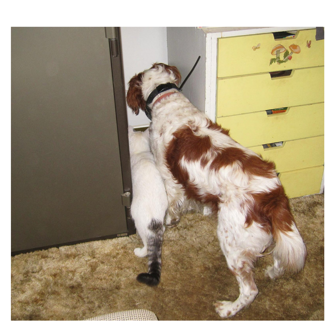
False Positive
(High n-gram similarity)

A computer *is sitting on a wooden table.*

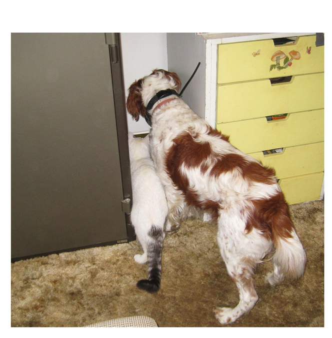


A large white flower *is sitting on a wooden table.*

False Negative
(Low n-gram and scene graph similarity)



A gray and white cat is standing next to a brown and white dog.



Two animals are looking at something in the wall.

OUR APPROACH

1. Build a text corpus from the list of captions and text descriptions accompanying the image descriptions in the dataset.
2. Generate a Latent Dirichlet Allocation (LDA) based Topic Model [1].
3. Train a Topic2Vec model by associating each caption with the appropriate topics [2,3].
4. Identify visual attributes using appropriate visual classifiers.
5. Derive topic distribution for given attributes using word-topic matrix generated from LDA.
6. Compute cosine similarity to quantify semantic relatedness of the caption to visual attributes present in the image.

$$y_i^a = \begin{cases} \text{Valid} & \text{if} \quad \cos\_sim(v_d^{c_i}, v_d^{a_i}) > \mathrm{T} \\ \text{Invalid} & \text{if} \quad \cos\_sim(v_d^{c_i}, v_d^{a_i}) \le \mathrm{T} \end{cases}$$

PRELIMINARY RESULTS

We show the visual similarity between sample caption from the MS COCO [5] dataset, and candidate visual attributes generated by an arbitrary visual classifier.

The most prominent visual concepts that are also described in some form in their respective captions have a higher similarity score.

Caption: "Two animals that are looking at something in the wall."

Visual Similarity: door (0.8022), dog (0.775), wall (0.7745), cat (0.7219), carpet (0.6552), cabinet (0.4725), cow (0.7515), horse (0.6816)

Caption: "A woman holding a clear umbrella in a dark city."

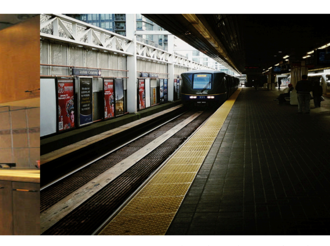
Visual Similarity: building (0.8574), crossing (0.6381), umbrella (0.6252), car (0.5635), woman (0.5580), cap (0.5506), man (0.5262), truck (0.5411)

Caption: "Some people wearing helmets are riding mopeds and some buildings."

Visual Similarity: bicycle (0.8618), motorcycles (0.7976), buildings (0.5452), woman (0.5574), grass (0.4407), man (0.3286), hat (0.6188), cow (0.5111)

Caption: "Dining room area with a stove and a small dining area in front of a window."

Visual Similarity: dining table (0.7706), chair (0.6168), stove (0.7189), window (0.6997), television (0.4349), table (0.4933)



Caption: "A train that is parked next to a train station."

Visual Similarity: train (0.9388), person (0.6521), bench (0.6002), television (0.5720), sky (0.6277), fence (0.5660)

Caption: "A table that has been serving a large tray of food."

Visual Similarity: vegetables (0.8768), meat (0.9018), drink (0.7986), man (0.7504), woman (0.7414), glass (0.6440), cap (0.5375), pizza (0.8841), clock (0.4503)

CONCLUSIONS AND FUTURE WORK

- Our approach successfully measures the degree of semantic relatedness between a natural language description and the visual attributes in an image in real-time.
- Does not require exhaustive set of reference captions.
- Also provides a form of weak-supervision labels for caption annotation in hybrid-intelligence systems.
- We would like to compare our approach with existing evaluation metrics such as SPICE.

REFERENCES

1. Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2002. Latent dirichlet allocation. In *Advances in neural information processing systems*, 601–608.
2. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J.; Sutskever, L.; and Zweig, G. 2014. word2vec.
3. Niu, L.; Dai, X.; Zhang, J.; and Chen, J. 2015. Topic2vec: learning distributed representations of topics. In *Asian Language Processing (IALP), 2015 International Conference on*, 193–196. IEEE.
4. Ramanathan, V.; Liang, P.; and Fei-Fei, L. 2013. Video event understanding using natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 905–912.
5. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

CONTACT

We're hiring and looking to collaborate with research groups. If you're interested, please reach out!

Divya Ramesh, Deep Learning Engineer - divya@cloudsight.ai
Brad Folkens, CEO - brad@cloudsight.ai