LT2318 - Artificial Intelligence: Cognitive Systems
Kate Viloria

"a close up of
a white background
with a black background"

When Your Model is a Bad Student

Course Project

Main Aim

Research Question:

How will a standard image captioning model perform in generating captions for the visually impaired?

ROUGH

Agenda

Purpose and Importance

Related Work

VizWiz Data and Model

Results & Discussion

Conclusion & Future Work

Why?

Real-World Application

- demand
- demographic
- direct effect

Scientific Contribution

- domain-specific task
- presents nuances that are not usually considered
- how a human would approach task

Related Work

Show and Tell

- neural image caption generator
- vision CNN → richer representations
- language generating RNN (LSTM-based)

Vinyals et al. (2015)

Neural Machine Translation

- proposes the attention mechanism
- look for parts in the source sentence that would be relevant to predicting target word

Bahdanau et al. (2014)

Show, Attend, and Tell

- neural image caption generation with visual attention
- prioritise salient features of an image
- low-level representation

Xu et al. (2015)

VizWiz Dataset

39,181 images taken by people who are visually impaired each with 5 human-made captions

VizWiz_train_00011794.jpg

- a black bottle with red and black label
- A bottle of hot sauce is full to the top.
- A bottle of brown liquid with a red and black label.
- Bottle of hot sauce sitting on what looks to be a pink pillow,
- A clear plastic bottle that is full of a green/brown colored liquid and has a red and black label with the word "HOT" on it and some pictures of palm trees.

Quick Facts

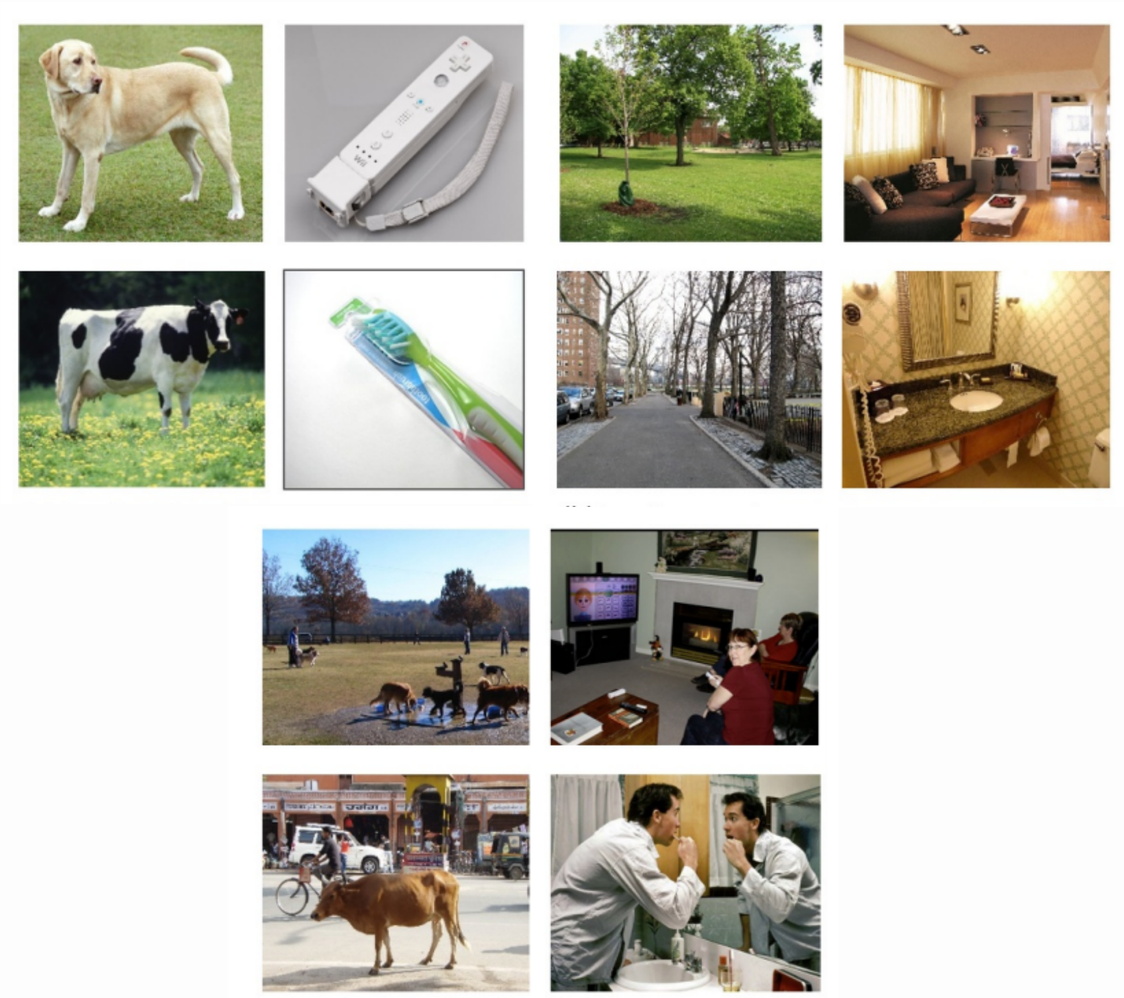
- domain-specific dataset
- image captioning, visual question answering, etc.
- Mechanical Turk (reliability and consistency)

Statistics

- training dataset: 23 431
- is_rejected
- is_precanned
- after filtering: 14 790
- 60/20/20 split

VizWiz
Dataset

MS-COCO

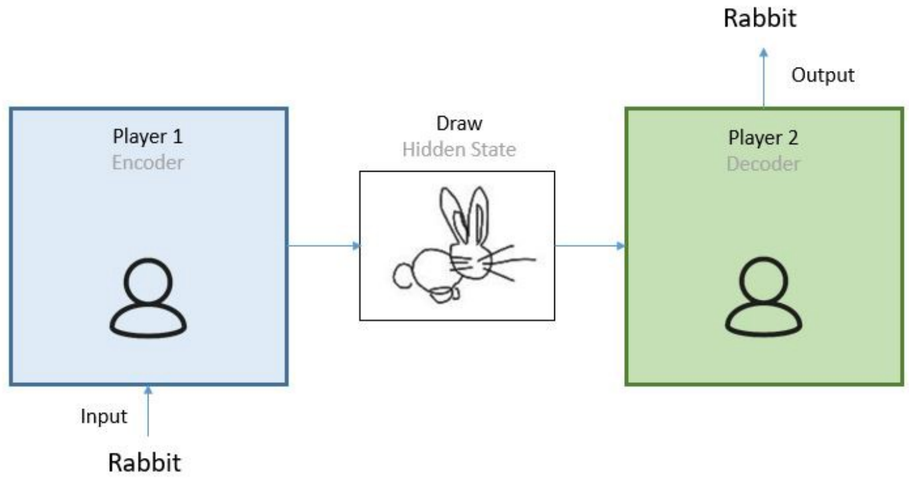


Lin et al. (2014)

Architecture

Model

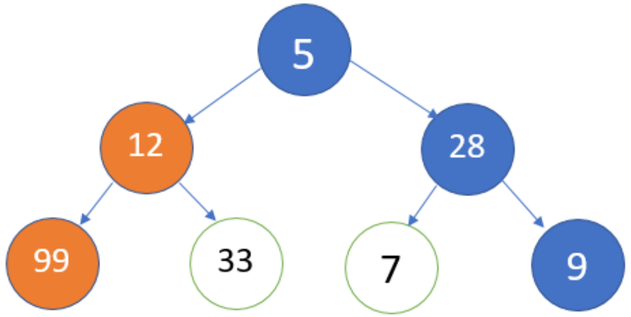
- encoder-decoder with attention mechanism
- pretrained encoder (ResNet)



https://towardsdatascience.com/what-is-an-encoder-decoder-model-86b3d57c5e1a

Algorithm

- greedy search for generating captions



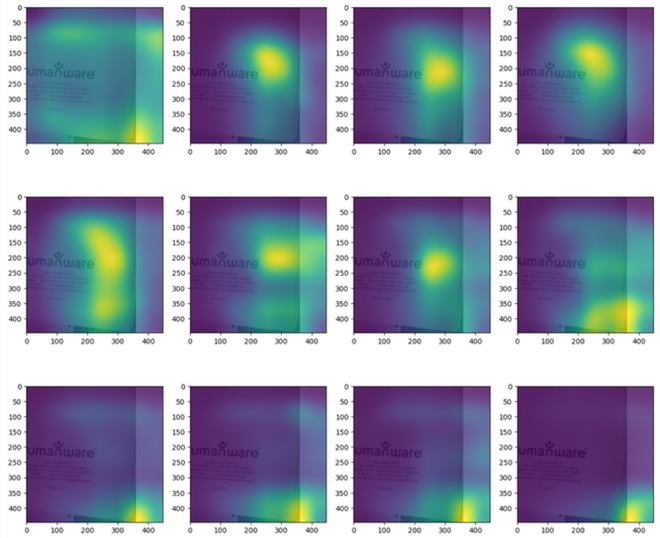
https://www.techopedia.com/definition/16931/greedy-algorithm

Other

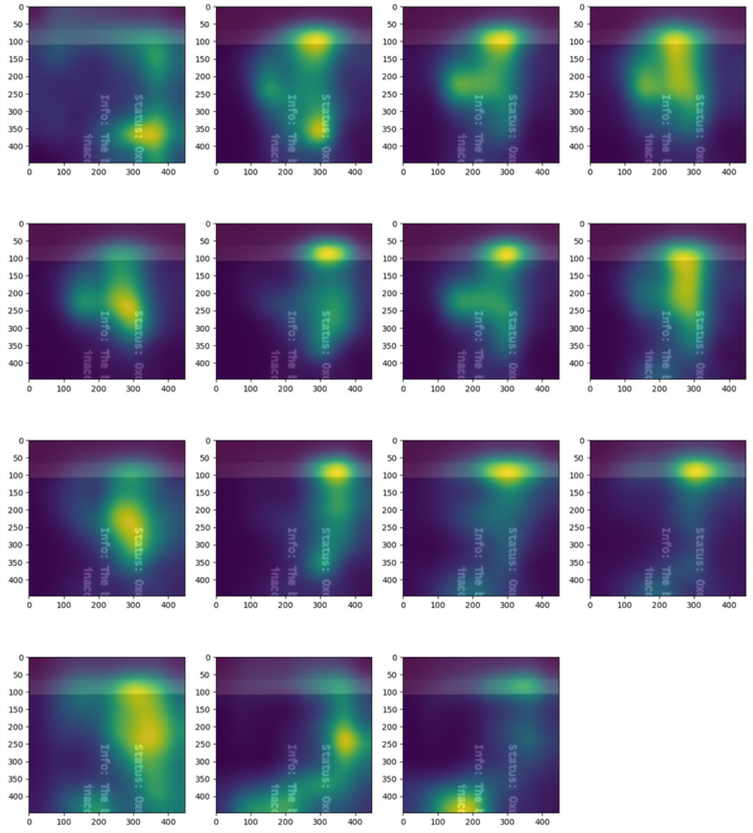
- padded sequences then ordered by length
- stop after 20 epochs of no improvement
- batch size = 32
- dim = 512 (for word embeddings, attention, and decoder)
- encoder LR = 0.0001
- decoder LR = 0.0004

Results

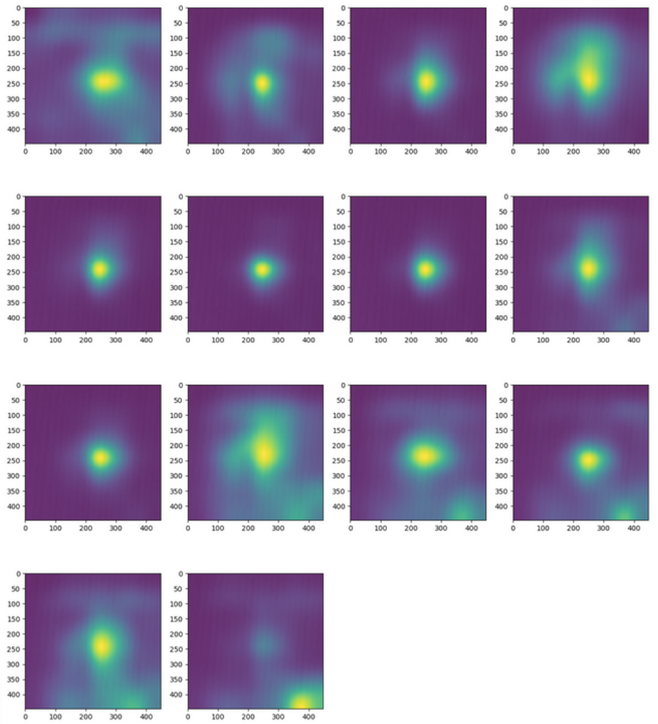
['a', 'close', 'up', 'of', 'a', 'white', 'background', 'with', 'a', 'black', 'background', '<end>']



['a', 'close', 'up', 'of', 'a', 'black', 'background', 'with', 'a', 'black', 'stripe', 'at', 'the', 'top', '<end>'] >']



['a', 'close', 'up', 'of', 'a', 'black', 'and', 'white', 'background', 'with', 'a', 'black', 'background', '<end>']



Results

It's Bad!!

Highest Checkpoint During Training
(Epoch 8)

LOSS = 4.114
TOP-5 ACCURACY = 64.060
BLEU-4 = 0.142178

Test Data

LOSS = 4.181
TOP-5 ACCURACY = 63.941
BLEU-4 = 0.145776

Corpus BLEU Score

- list of references (human-made annotations)
- candidate (predicated caption)
- average the scores between the candidate and each reference
- longer sentences lead to lower scores

The captions generated...

- were mostly the same caption but with some variations
- appeared to be recursive
- were uninfomative
- were just...bad!

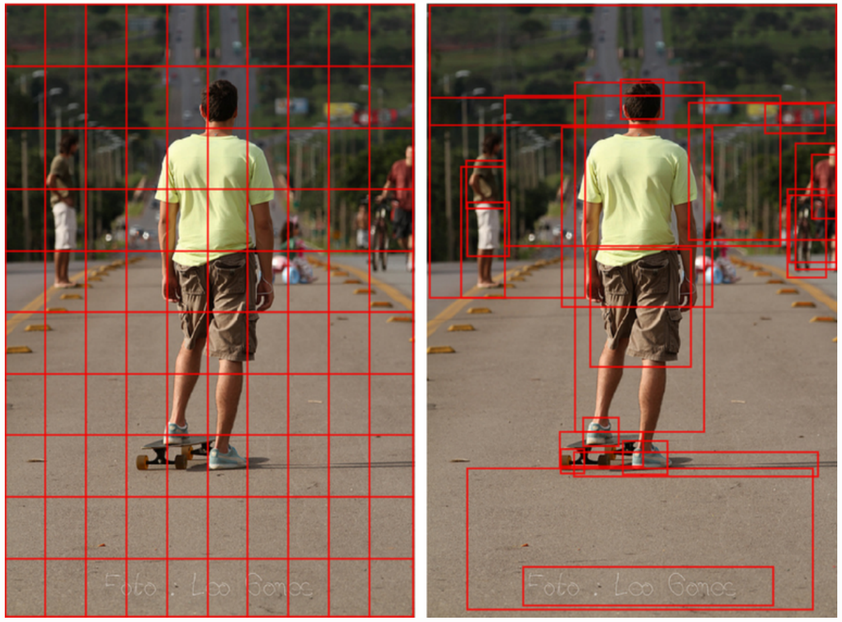
Discussion

Why so bad?????

A LOT OF REASONS

- domain-specific
- human-annotated captions have a common template and structure but also extremely varied
  - foreground, any text, background
  - a bad combination with greedy search!
- a standard model
- top-down instead of bottom-down approach
  - objects usually partial in images

Top-Down **Bottom-Up**



Anderson et al. (2018)

Conclusion & Future Work

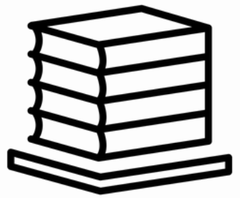
- shortcomings of algorithms and overlooked limitations of data will yield bad results!
- must be critical of dataset being used!
  - instructions to annotators must be considered
- model is simply too simple!

FUTURE WORK

- beam search
- visual sentinel proposed by Lu et al. (2017)
- an initial rule-based template for captions
- DenseCap presented by Johnson et al. (2016)
  - generates phrases over regions of an image
- optical character recognition (OCR)
- text summarization

Thank you! 🙂

References



Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Gurari, D., Zhao, Y., Zhang, M., & Bhattacharya, N. (2020, August). Captioning images taken by people who are blind. In European Conference on Computer Vision (pp. 417-434). Springer, Cham.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

Images by Appearance
bad grades by Gan Khoon Lay from the Noun Project
goal by Vectors Point from the Noun Project
work time by Bernar Novalyi from the Noun Project
helping hand by Jonathan Meyer from the Noun Project
Educational Background by Massupa Kaewgahya from the Noun Project
Architecture by Flatart from the Noun Project
Results by monkik from the Noun Project
discuss by Bernar Novalyi from the Noun Project
improve by Adrien Coquet from the Noun Project
Smile by abderraouf omara from the Noun Project
books by priyanka from the Noun Project