Visual Question Answering via Dense Captioning

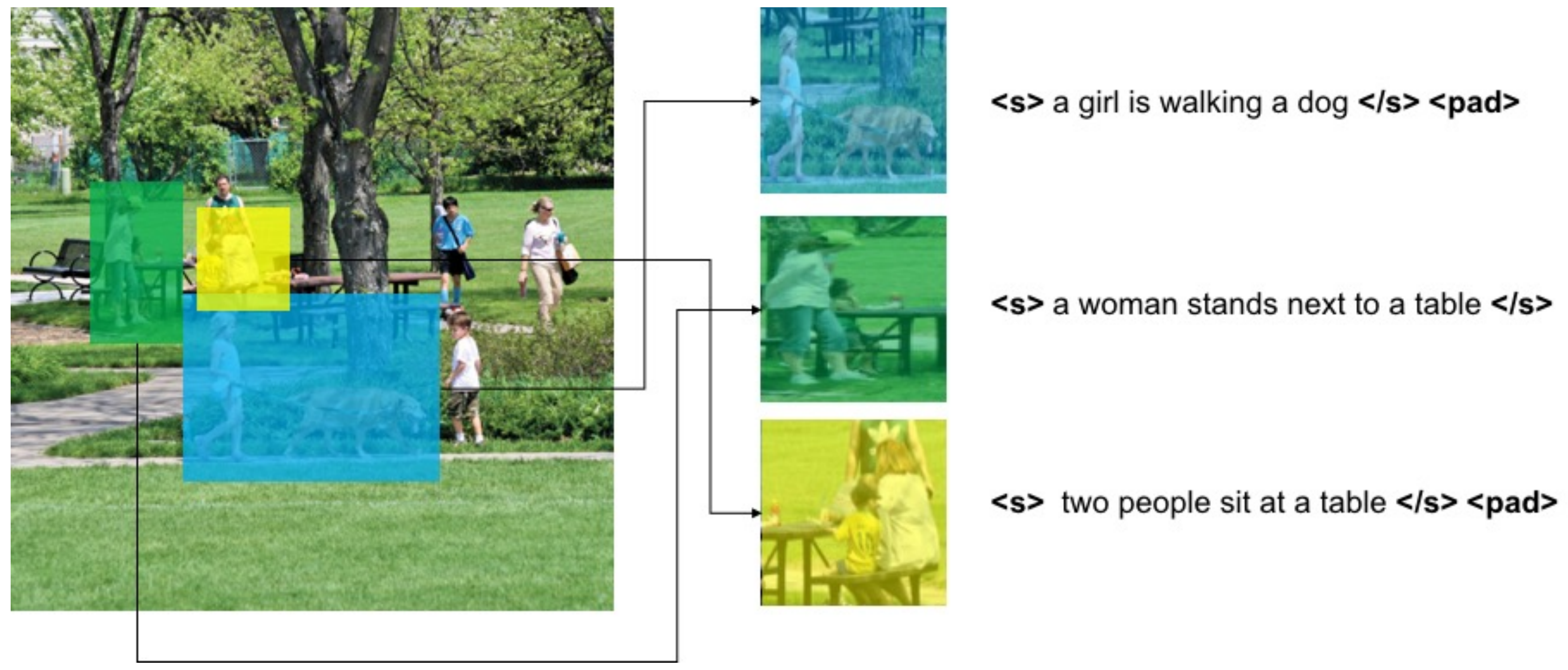
Neel Ramachandran, Emmie Kehoe, Vinay Sriram

Stanford University

PROBLEM DEFINITION

- Historical approaches to visual question answering (VQA) have relied on end-to-end models trained using a corpus of only images coupled with associated {Question, Answer} pairs.
- Our approach instead performs answering in two stages: (1) dense captioning to produce passages of region descriptions and (2) pure text question answering on generated passages.
- We use the Visual Genome Dataset's region descriptions and question answer pairs to train the two stages of our model.

REGION CAPTION PREPROCESSING

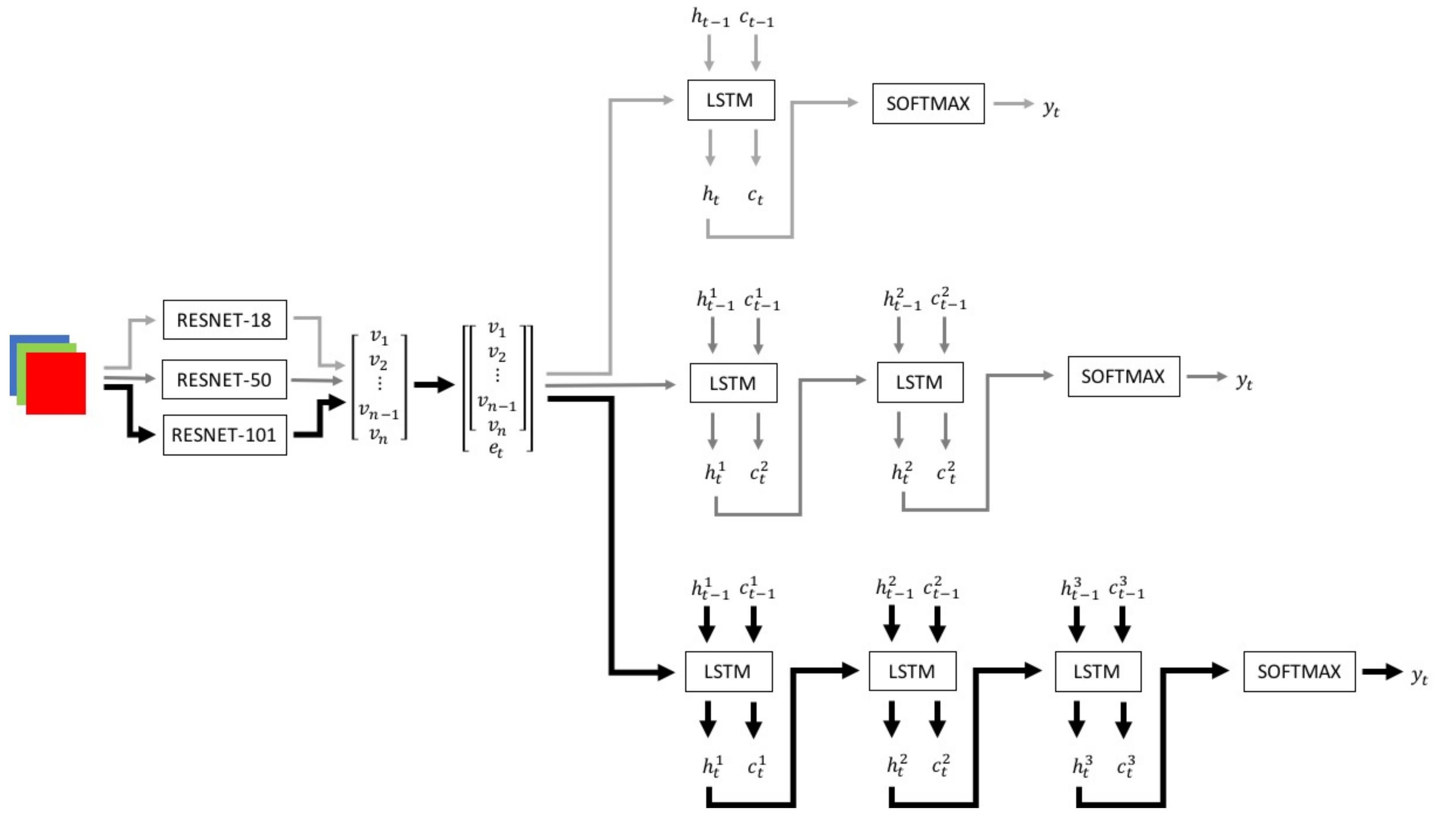


<s> a girl is walking a dog **</s> <pad>**

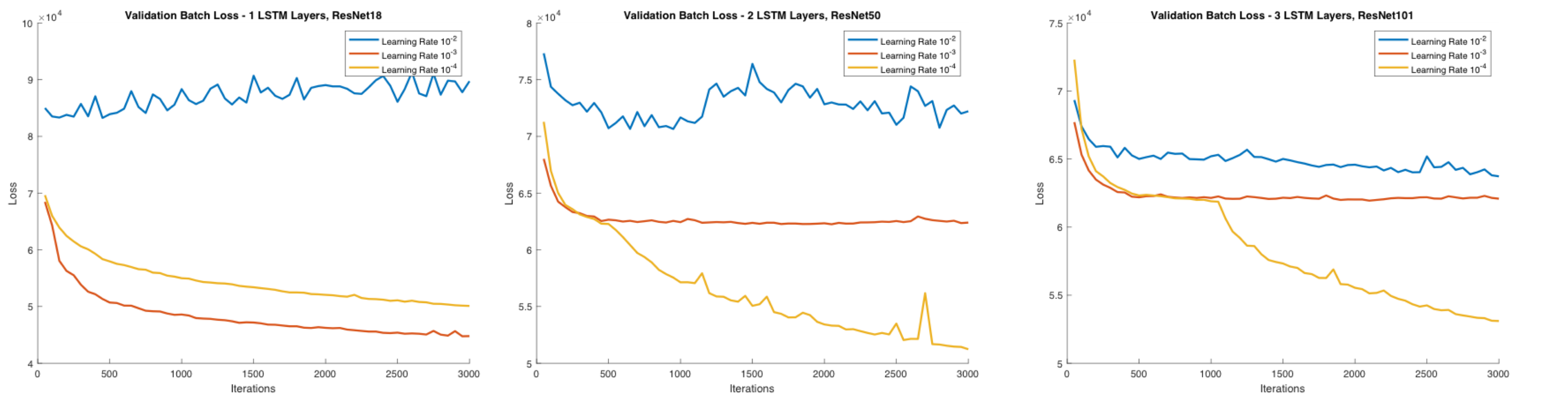
<s> a woman stands next to a table **</s>**

<s> two people sit at a table **</s> <pad>**

CAPTIONING MODELS



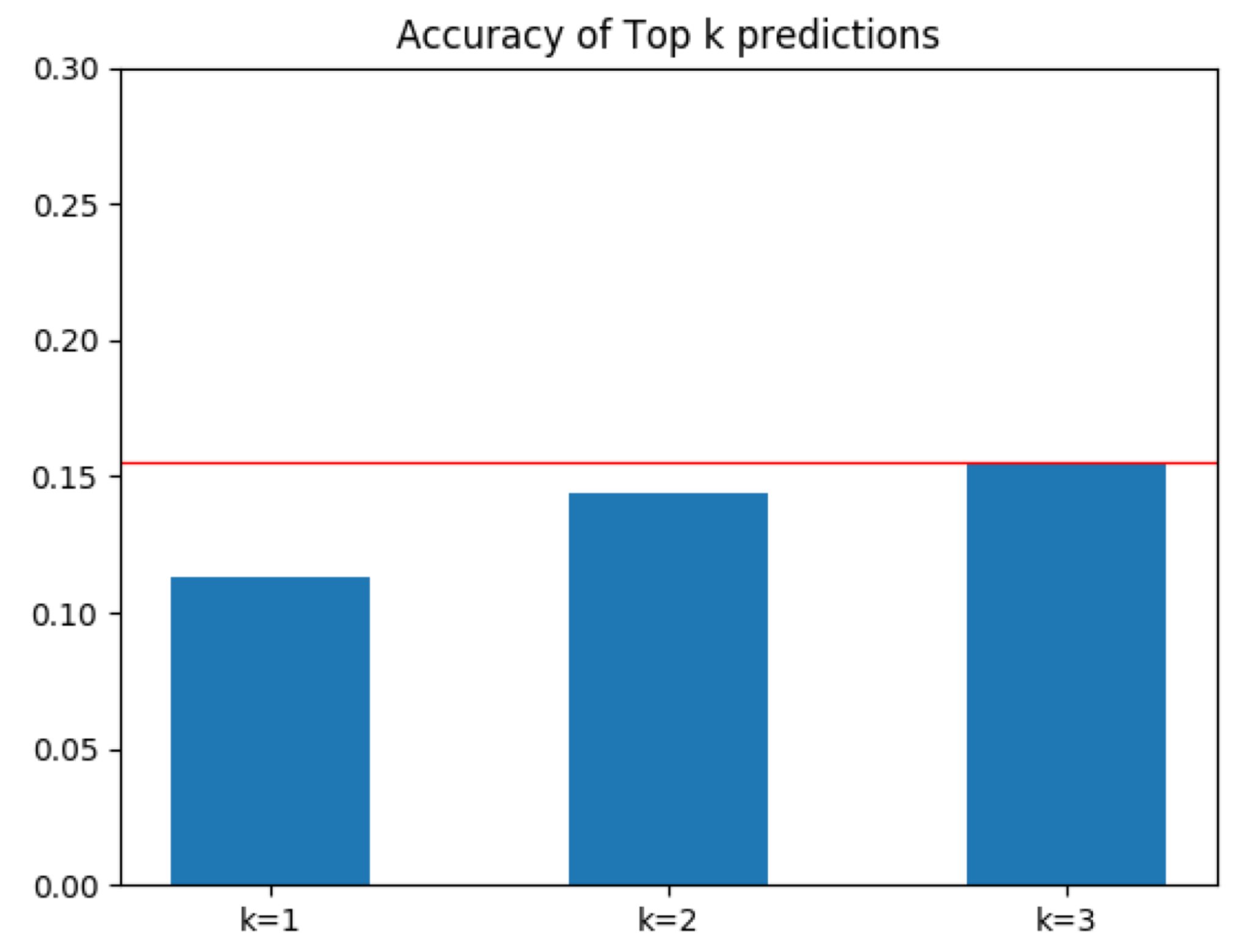
EXPERIMENTS



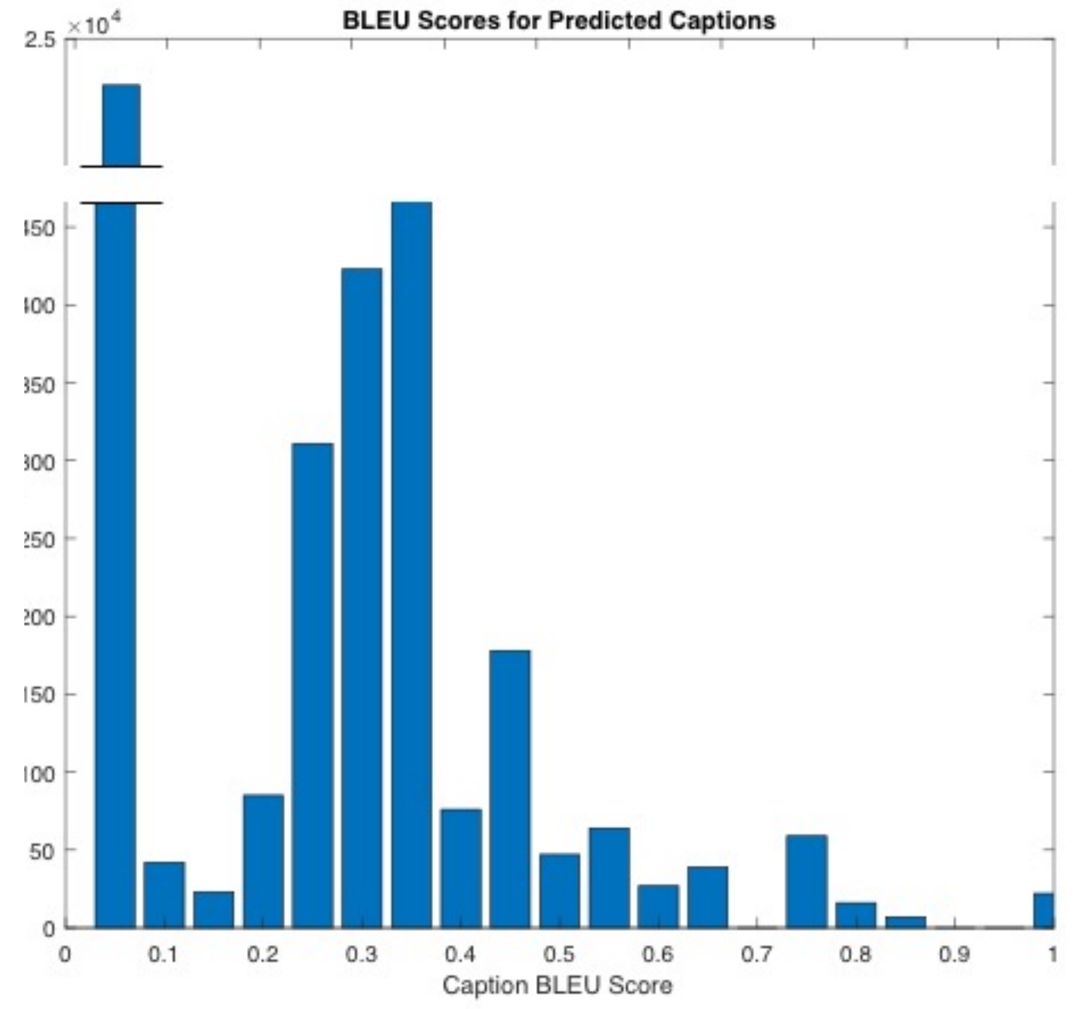
Captioning Model: From 200,000 generated bounding box region pairs, we produce a train:test:val split with ratio 60,000:20,000:20,000. For periodic evaluation, we randomly sample the validation set by 10x for faster speed. Hyperparameter tuning identifies the ResNET-101 architecture as best-performing. We select this model and train for an extended period (50,000 iterations) using a learning rate of 10^{-4}.

Question-Answering Model: Once we find region captions for an image, we reduce the visual question-answering task to a purely text-based question-answering task. We consider only question-answer pairs with one-word answers and therefore run a simplified Bidirectional Attention Flow (BiDAF) model that predicts a single index. After training the simplified BiDAF model on ground-truth region descriptions, we are able to achieve 60% exact-match accuracy on the validation set.

QUANTITATIVE RESULTS

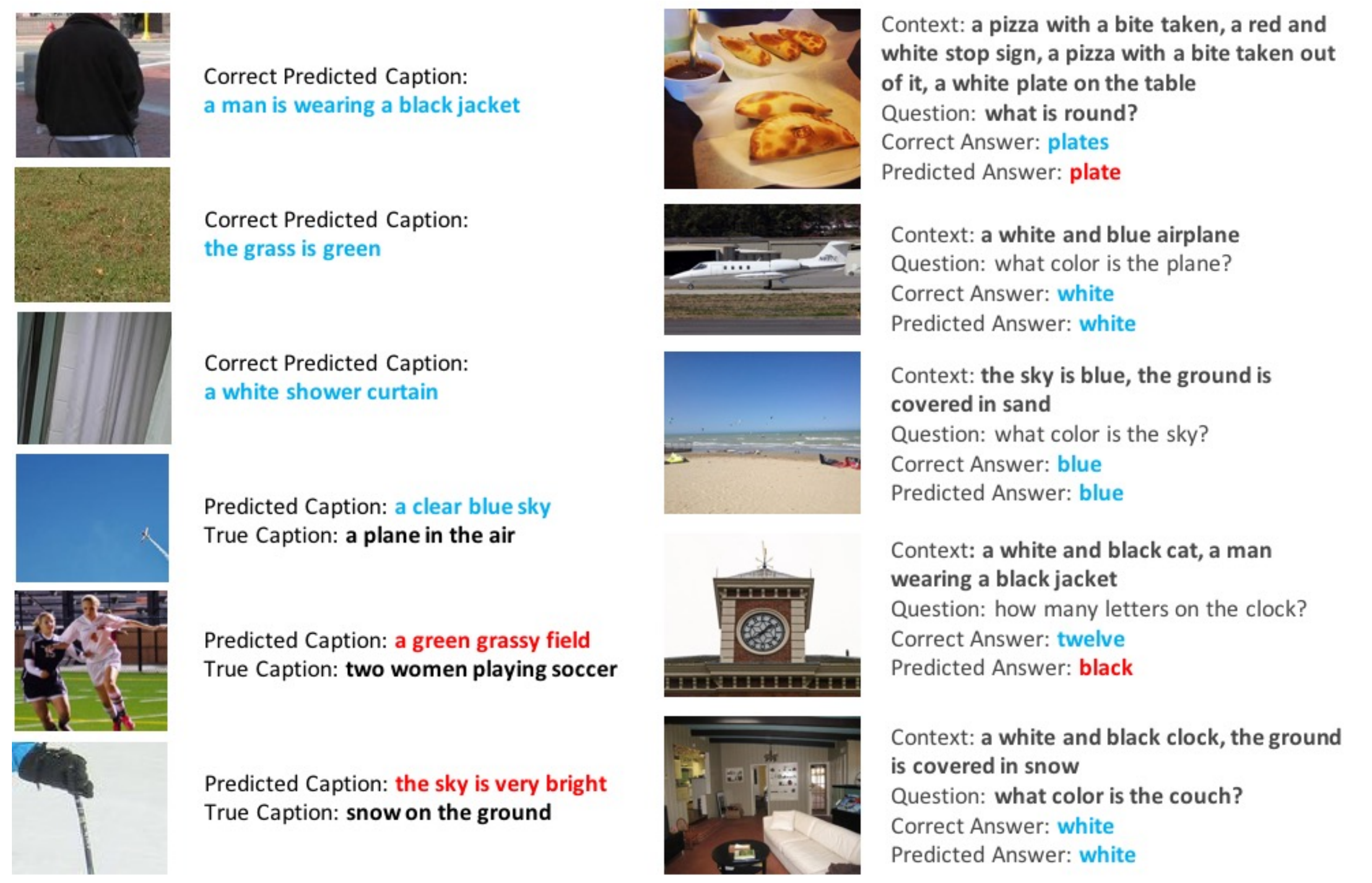


Question Answering Accuracy



Captions Test Set BLEU Scores

QUALITATIVE RESULTS



Correct Predicted Caption:
a man is wearing a black jacket

Correct Predicted Caption:
the grass is green

Correct Predicted Caption:
a white shower curtain

Predicted Caption: a clear blue sky
True Caption: **a plane in the air**

Predicted Caption: a green grassy field
True Caption: **two women playing soccer**

Predicted Caption: the sky is very bright
True Caption: **snow on the ground**

Context: **a pizza with a bite taken, a red and white stop sign, a pizza with a bite taken out of it, a white plate on the table**
Question: **what is round?**
Correct Answer: plates
Predicted Answer: plate

Context: **a white and blue airplane**
Question: what color is the plane?
Correct Answer: white
Predicted Answer: white

Context: **the sky is blue, the ground is covered in sand**
Question: what color is the sky?
Correct Answer: blue
Predicted Answer: blue

Context: **a white and black cat, a man wearing a black jacket**
Question: how many letters on the clock?
Correct Answer: twelve
Predicted Answer: black

Context: **a white and black clock, the ground is covered in snow**
Question: **what color is the couch?**
Correct Answer: white
Predicted Answer: white

ANALYSIS & DISCUSSION

- 15% of our captions contain the answer word, which provides an upper bound on our overall two-stage question-answering accuracy.
- We achieve a single-guess accuracy of 11%, and are able to nearly reach the upper bound by considering our top-3 guesses per question.
- Our model performs well on questions that ask about explicit features of the image (e.g. 25% accuracy on "What color?" questions), and poorly on more abstract questions whose answers are not well-captured by descriptions of individual regions within the image (e.g. 0% accuracy on "What time of day?" questions).
- With improvements to both stages, we see the multi-stage model as an effective way to provide greater interpretability to VQA results.

REFERENCES

Anderson, Peter et al. ``Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

Krishna, Ranjay, et al. ``Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." International Journal of Computer Vision, vol. 123, no. 1, 2017, pp. 32–73., doi:10.1007/s11263-016-0981-7.