End-to-end Image Captioning Exploits Multimodal Distributional Similarity

Pranava Madhyastha
p.madhyastha@sheffield.ac.uk
Josiah Wang
j.k.wang@sheffield.ac.uk
Lucia Specia
l.specia@sheffield.ac.uk

Department of Computer Science
The University of Sheffield
Sheffield, UK

A Hyperparameter Settings

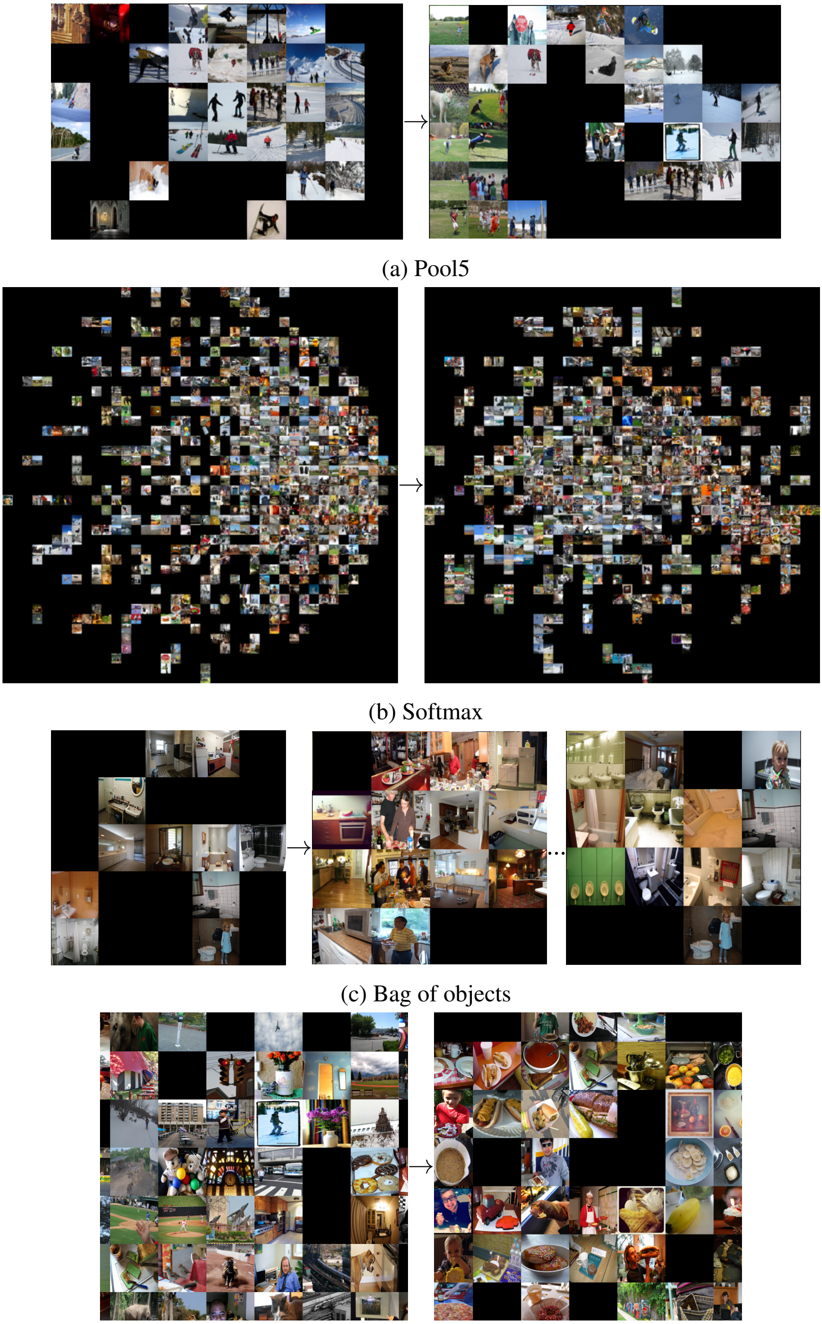
Our model settings were:

- LSTM with 128 dimensional word embeddings and 256 dimensional hidden representations Dropout over LSTM of 0.8
- We used Adam for optimization.
- We fixed the learning rate to 4e-4

We report our results by keeping the above settings constant.

B Analyzing Transformed Image Representations: Enlarged Figures

Figure 1 shows an enlarged version of Figure 1 in the main paper for better viewing.



(a) Pool5

(b) Softmax

(c) Bag of objects

(d) Pseudo-random

Figure 1: Visualization of the t-SNE projection of initial representational space (left) vs. the transformed representational space (right). See main text for a more detailed discussion. Please find the original images here: https://github.com/sheffieldnlp/whatIC