End-to-end Image Captioning Exploits Multimodal Distributional Similarity

Pranava Madhyastha, Josiah Wang, Lucia Specia

Department of Computer Science, University of Sheffield, UK

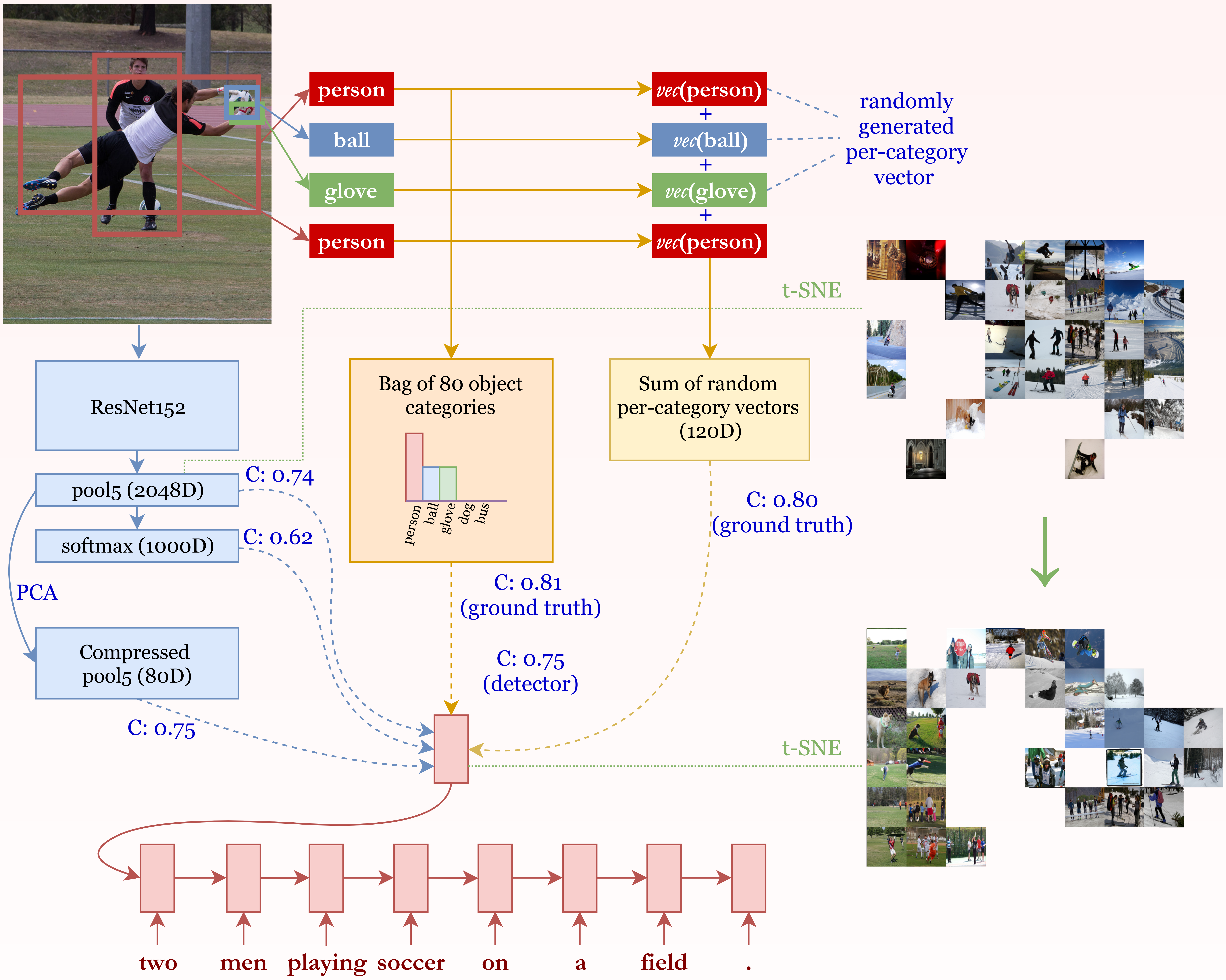
Research question

- Does end-to-end image captioning systems really exploit as much image information as people think? We say no.
- “You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

Hypothesis

- End-to-end image captioning systems exploit:
  - *distributional similarity* (image matching)
  - in a projected *multimodal* feature space.
- They generate a caption from similar examples in the training set.

Experiments to investigate hypothesis: Keep language model constant, vary image representations



Take home messages

- End-to-end image captioning models perform image retrieval, not image understanding.
- End-to-end image captioning models learn a joint textual-visual semantic subspace.
- End-to-end image captioning models can separate structure from noise.
- End-to-end image captioning models suffer no significant losses when the image representation is factorized to a low-dimensional space.
- There is scope to exploit more from images than is currently done.