Towards Unsupervised Image Captioning with Shared Multimodal Embeddings

Iro Laina¹  Christian Rupprecht²  Nassir Navab¹

Overview

We explore unsupervised image captioning, where image and language sources are independent (no image-caption pairs).

Language Domain

Training in two steps:
- Language model on text corpus only to learn a sentence embedding.
- Align images with the same embedding space, such that images and sentences are indistinguishable for the decoder.
- Our method allows to freely combine image and text sources for a wide variety of image types and text styles.
- Minimal supervision through an off-the-shelf object detector to perform an initial domain alignment.

Language Domain

Training objective:  \( L_{CE}(\mathbf{e}) + \lambda_1 L_{triple}(\mathbf{e}, \mathbf{e}, \mathbf{e}) \)

Visual similarity in language domain:

\( \mathbf{e} = \text{man.n.01, bicycle.n.01, road.n.01} \)

Two men sitting on a bench with their bikes parked on the side

\( \geq 2 \)

Domain Alignment

1. Find weak correspondences between language and image domains
2. Train with assigned weak labels as supervision

Ablation Experiments (Unpaired COCO)

COCO Images / Conceptual Captions

- Little boy playing with a teddy bear
- Young woman holding a glass of wine
- Young man working on a laptop in the office

COCO Images / VQA-v2

- Is the teddy bear wearing glasses?
- Is the man on the bench wearing headphones?

COCO Images / J.R.R. Tolkien books

- Giraffe eating leaves from a tree
- I know what you mean. There might be all the difference between an old cow sitting and thoughtfully chewing, and an 'unk', and the change might come suddenly.

Quantitative Evaluation

under the unpaired setting (COCO)

Feng et al.[2] 18.6 17.9 54.9 11.1 ours 19.3 20.2 61.8 12.9

under the unsupervised setting (COCO/GCC)

Ours 8.3 13.3 28.9 7.2

References