Captioning Images with Diverse Objects.

Subhashini Venugopalan[1], Lisa Anne Hendricks[2], Marcus Rohrbach[2,3], Raymond Mooney[1], Trevor Darrell[2], Kate Saenko[4]

[1] UT-Austin [2] UC-Berkeley [3] Facebook AI Research [4] Boston University

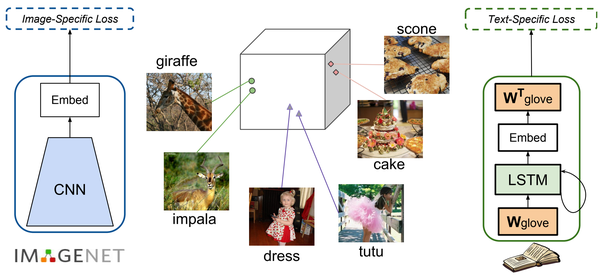
Goals

Existing visual classifiers can recognize hundreds of categories of objects. Can we describe these objects in context without paired image-caption training data?

We propose **N**ovel **O**bject **C**aptioner which can describe objects unseen in paired image-caption data.

NOC Key Insights

Train jointly on multiple data sources.

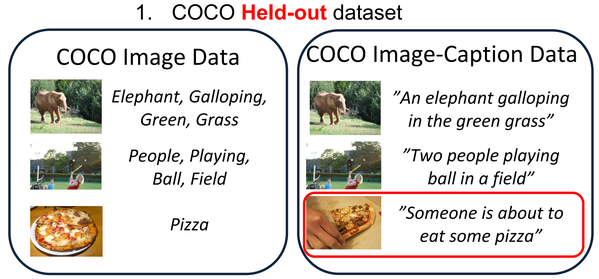


1. **Learn from unpaired data.** Train visual CNN on unpaired image data, and an LSTM Language Model on unannotated text data.
2. **Capture semantic similarity** of words in the language model using dense word embeddings.
3. **Train jointly to describe novel objects.** A visual recognition CNN, a language model, and an image-caption model [1] are trained jointly on different data sources with shared parameters.

Evaluation

We hold out a subset of data from COCO [2].

1. COCO **Held-out** dataset

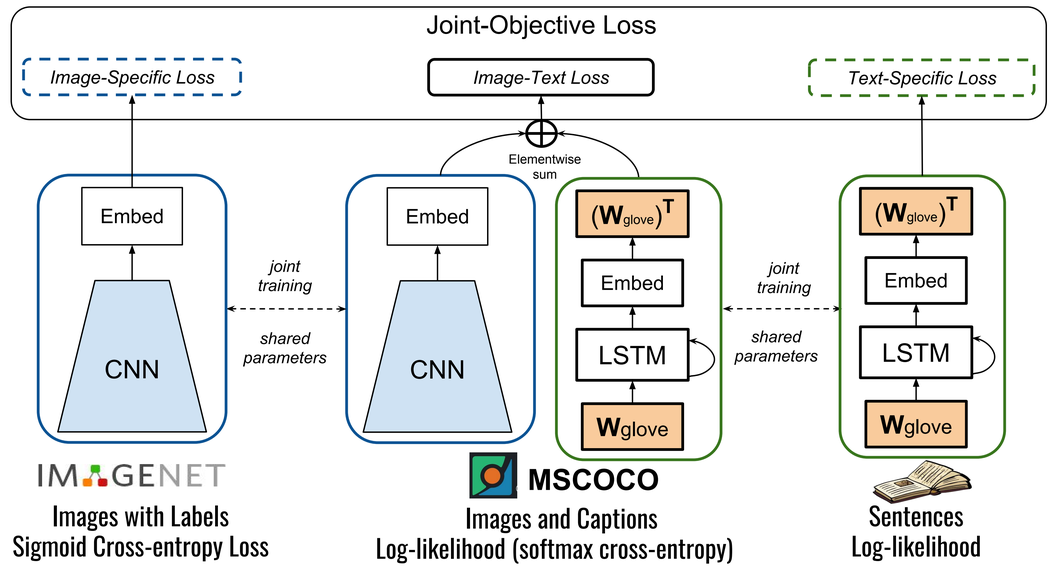


2. IMAGENET

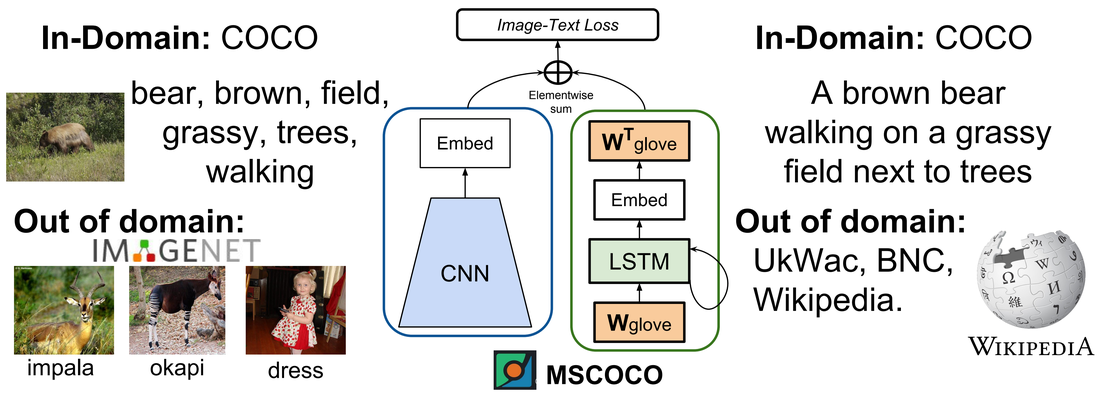
638 categories from ImageNet not mentioned in COCO. 52 classes with rare mentions (med∼5 images) in COCO.

Model

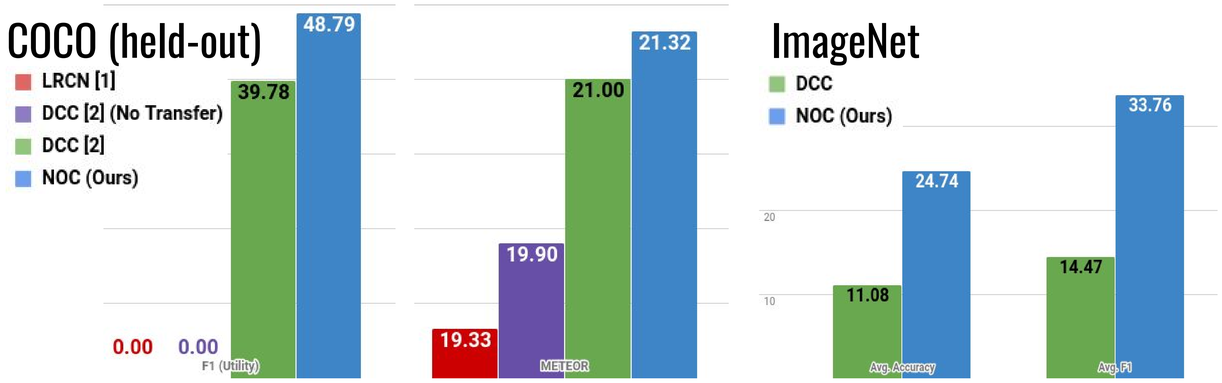
Share network parameters and train jointly on multiple data sources and with different objectives.



Training Data



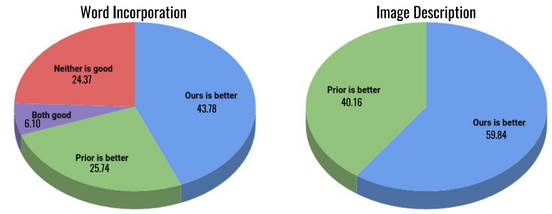
Results



F1 (Utility): Ability to recognize and incorporate new words. **METEOR**: Fluency and sentece quality.

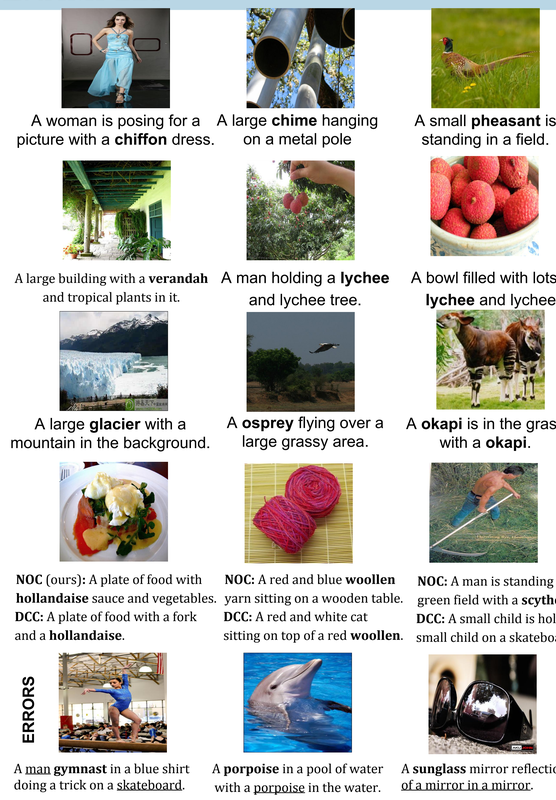
ImageNet Human Eval.

Word Incorporation: Which model incorporates the word (name of the object) in the sentence better?
Image Description: Which describes the image better?



Intersection (both DCC and NOC can caption): NOC maintains descriptive quality but captions more objects.

Examples



Code and References

Project Page
http://vsubhashini.github.io/noc.html

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[2] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.