Captioning Images with Diverse Objects


Subhashini Venugopalan


Lisa Anne Hendricks


Marcus Rohrbach


Raymond Mooney


Kate Saenko

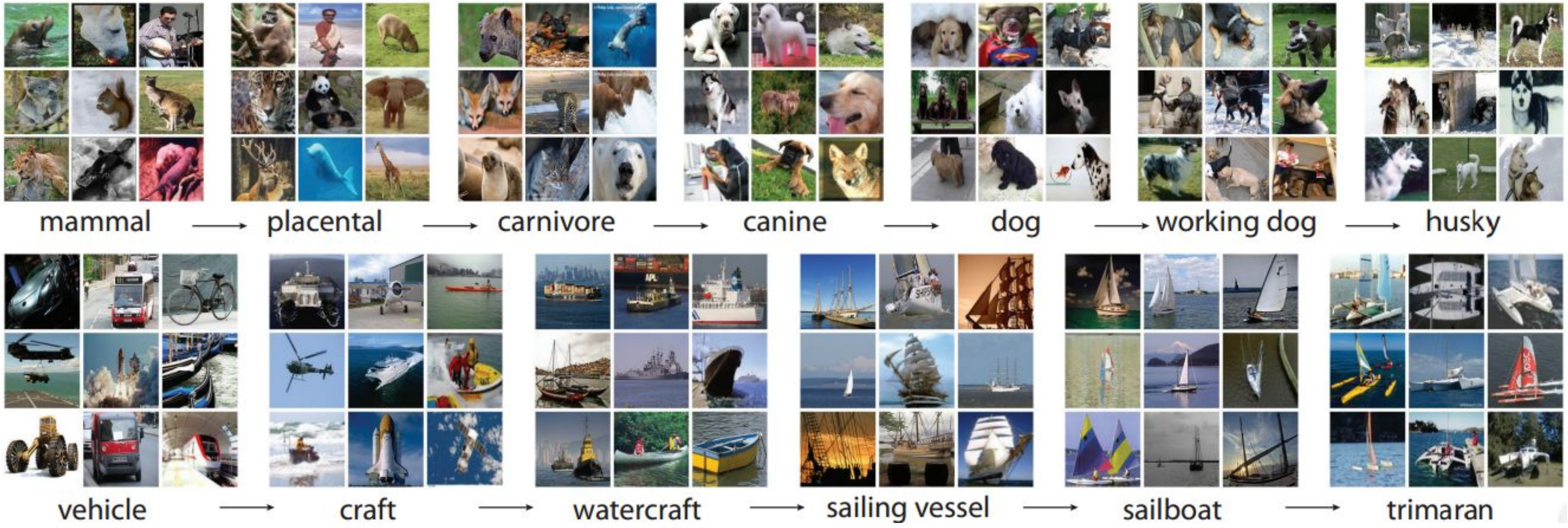

Trevor Darrell

UT Austin UC Berkeley Boston Univ.

Object Recognition

Can identify hundreds of categories of objects.

IMAGENET 14M images, 22K classes [Deng et al. CVPR'09]



Visual Description

Berkeley LRCN [Donahue et al. CVPR'15]:
A brown bear standing on top of a lush green field.

MSR CaptionBot [http://captionbot.ai/]:
A large brown bear walking through a forest.

MSCOCO
80 classes

Novel Object Captioner (NOC)

We present Novel Object Captioner which can compose descriptions of 100s of objects in context.

NOC (ours): Describe novel objects without paired image-caption data.

IMAGENET + MSCOCO +

→ An **okapi** standing in the middle of a field.

Visual Classifiers. IMAGENET

→ **okapi**

Existing captioners. IMAGENET init + train MSCOCO

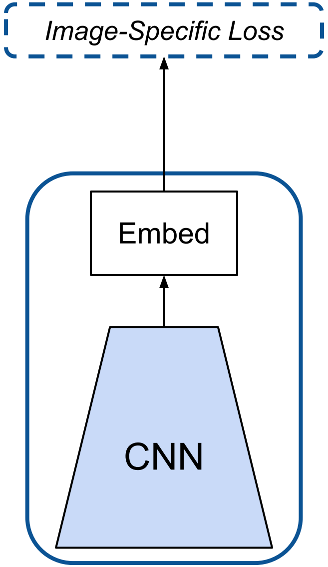
→ A horse standing in the dirt.

Insights

1. Need to recognize and describe objects outside of image-caption datasets.

okapi

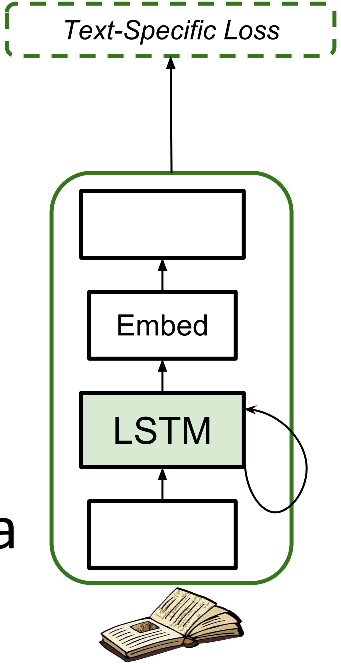
Insight 1: Train effectively on external sources



IMAGENET

Visual features from unpaired image data

Language model from unannotated text data



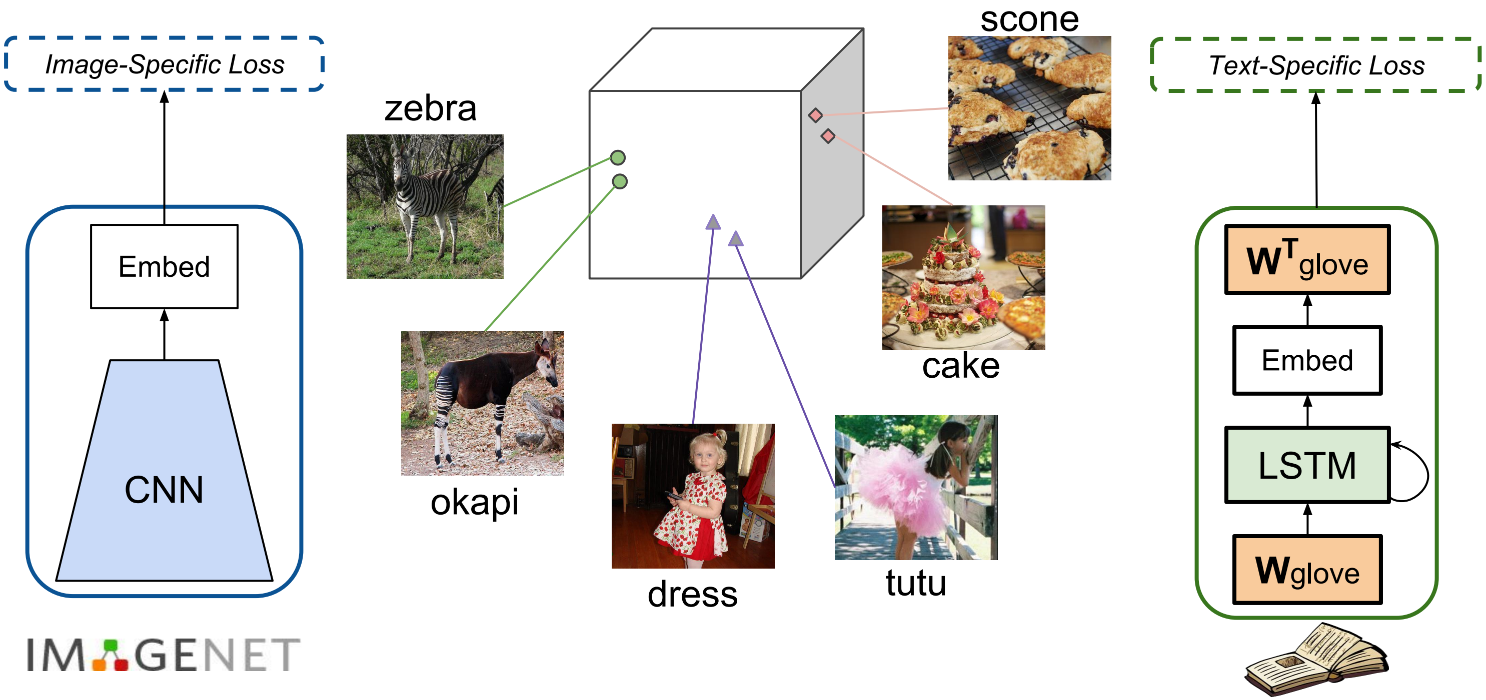
Insights

2. Describe unseen objects that are similar to objects seen in image-caption datasets.

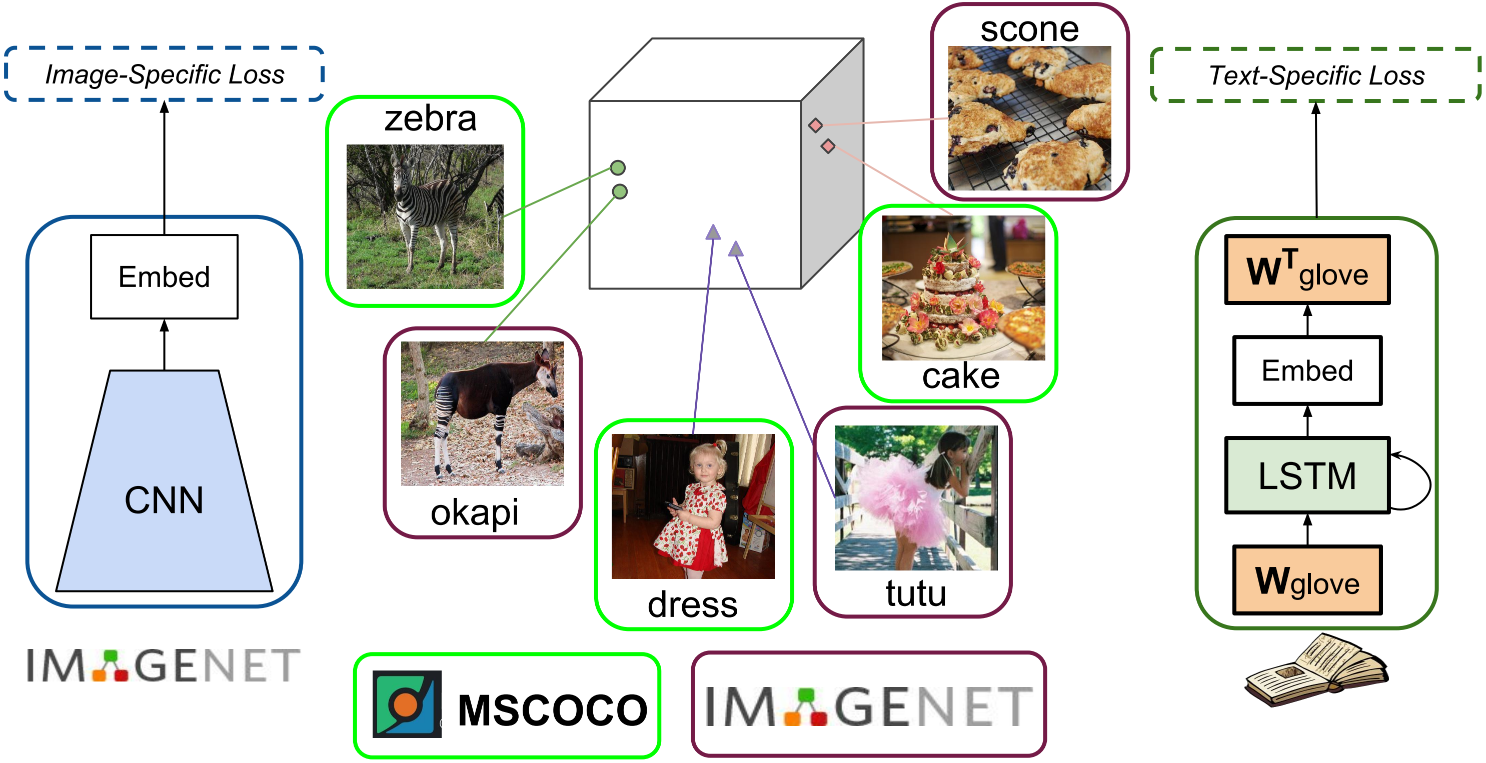
okapi

zebra

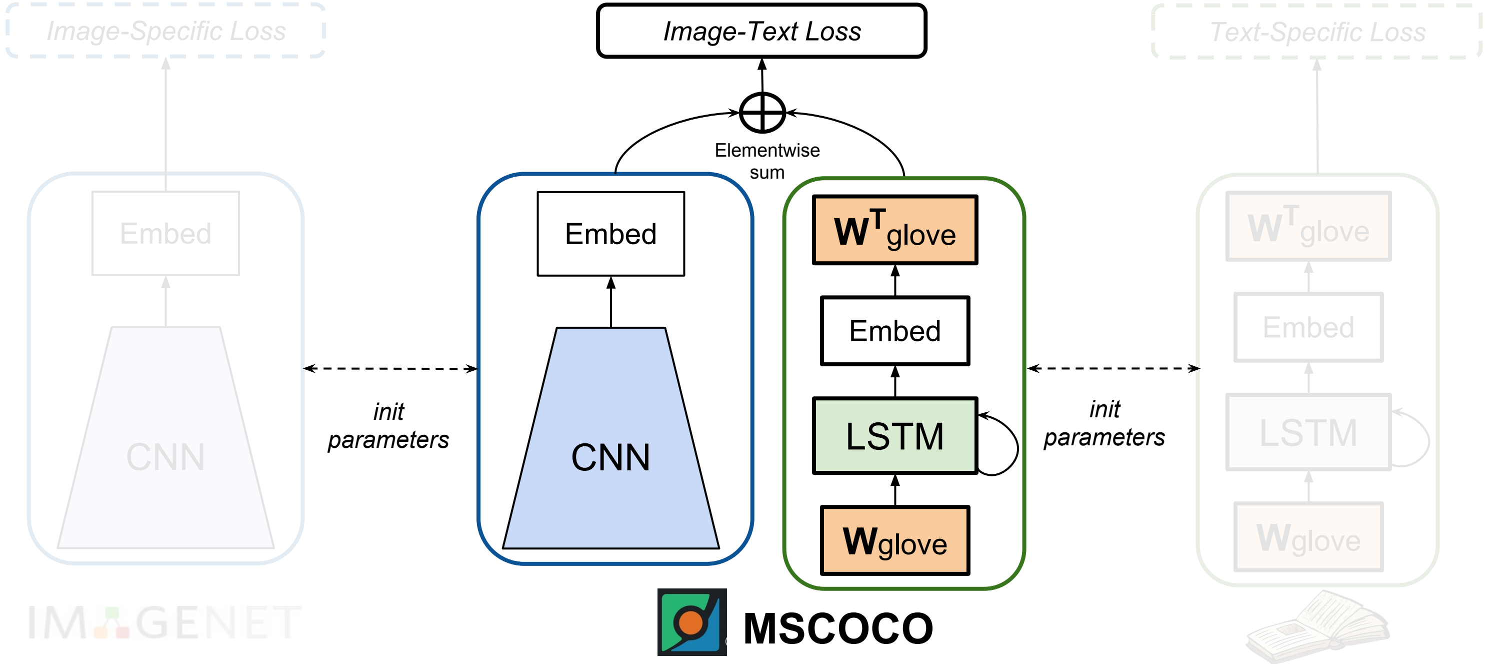
Insight 2: Capture semantic similarity of words



Insight 2: Capture semantic similarity of words



Combine to form a Caption Model



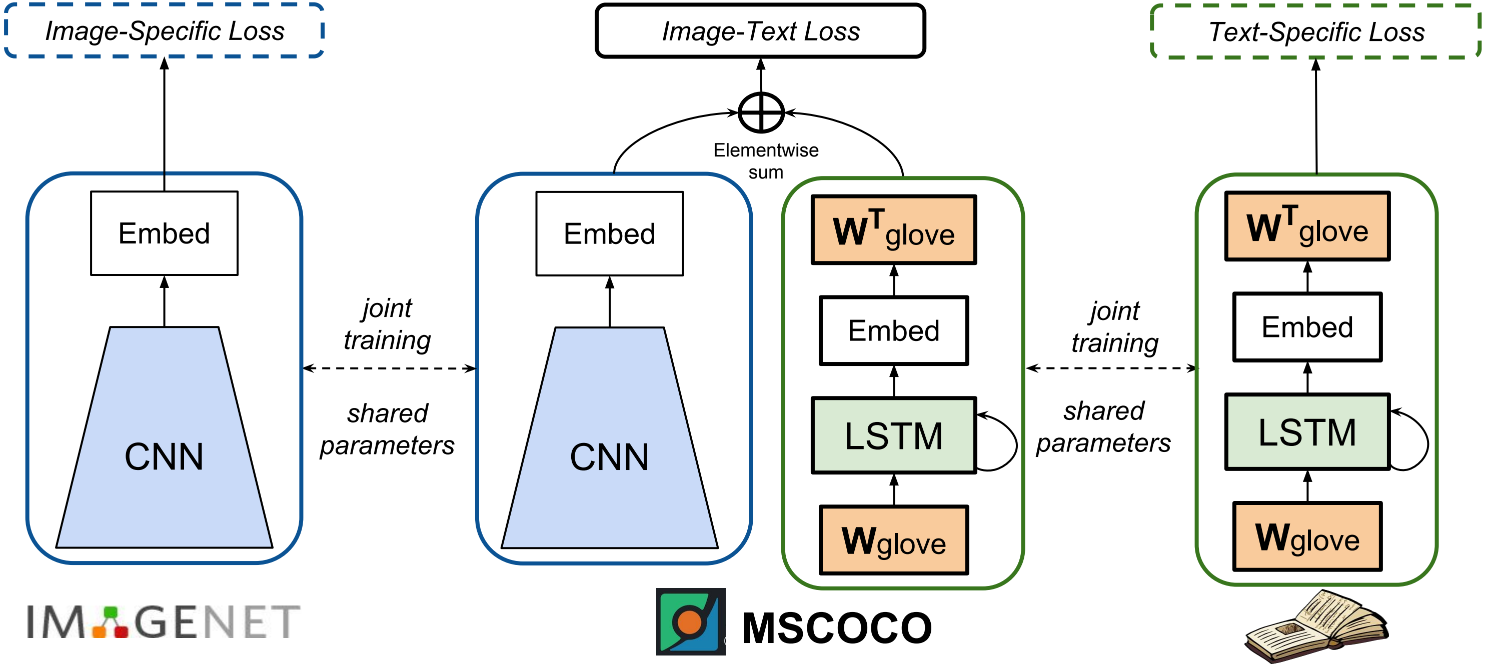
Not different from existing caption models. Problem: Forgetting.

Insights

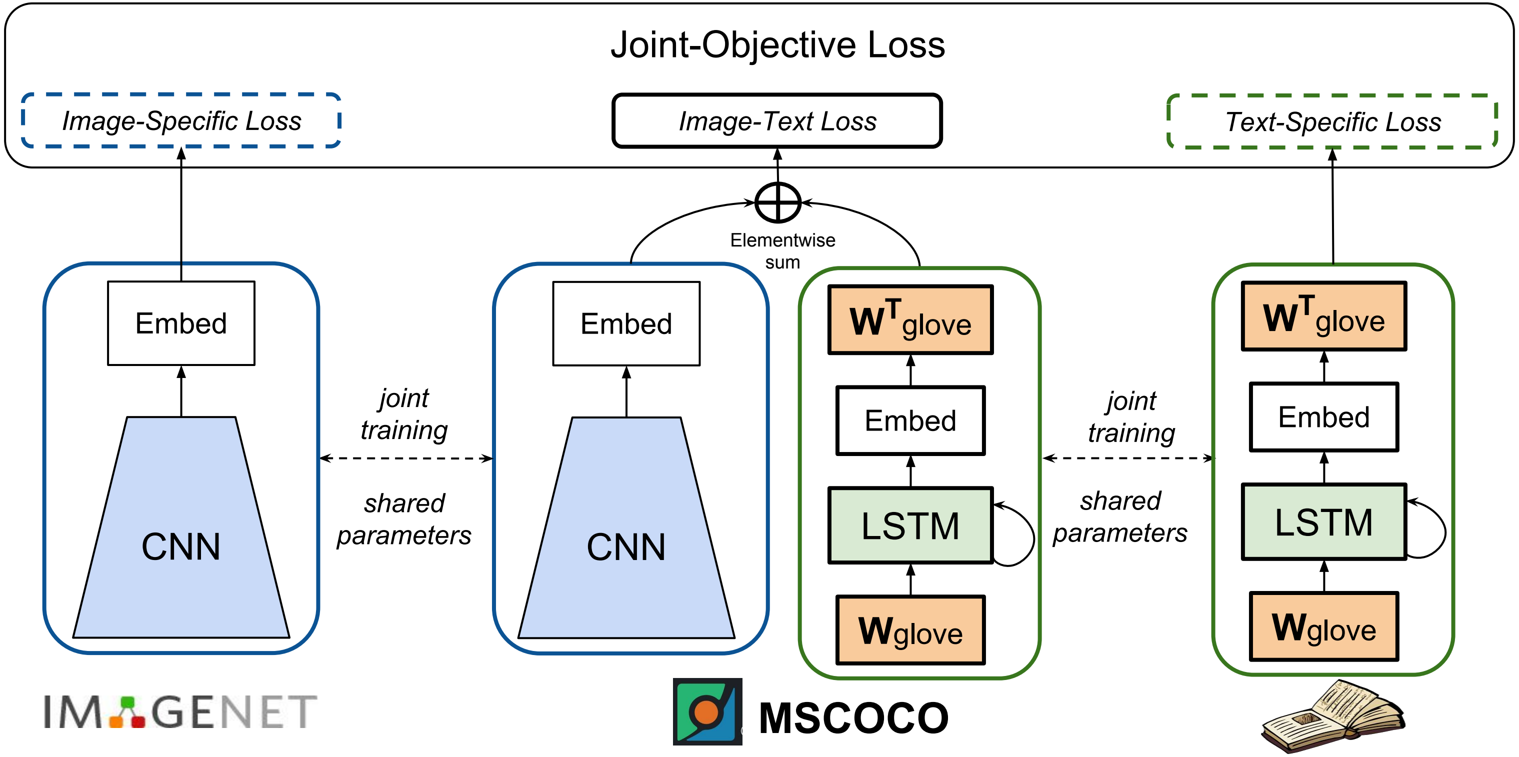
3. Overcome “forgetting” since pre-training alone is not sufficient.

[Catastrophic Forgetting in Neural Networks. Kirkpatrick et al. PNAS 2017]

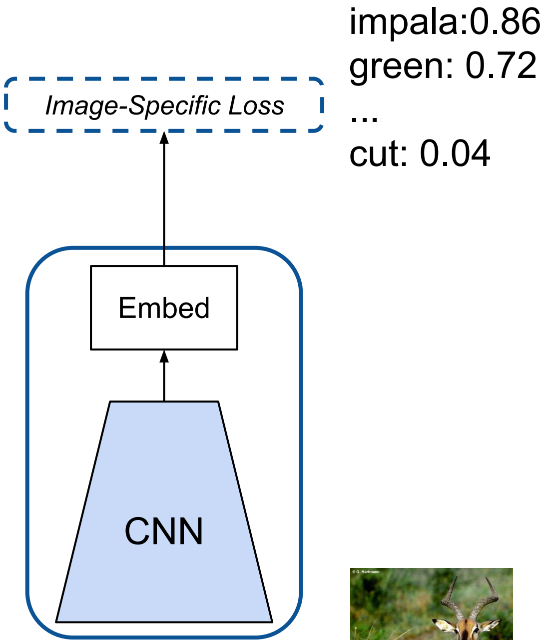
Insight 3: Jointly train on multiple sources



Novel Object Captioner (NOC) Model



Visual Network



Network: VGG-16 with multi-label loss [sigmoid cross-entropy loss]

Training Data: Unpaired image data

Output: Vector with activations corresponding to scores for *words in the vocabulary*.

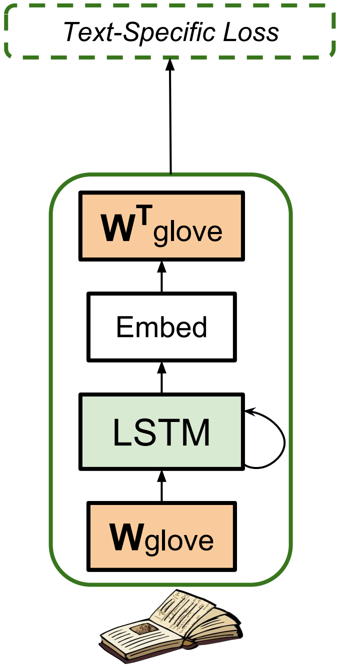
Language Model

Network: Single LSTM layer. Predict next word w_{t+1} given previous words $w_{0..t}$ $p(w_{t+1} \mid w_{0..t})$

$(\mathbf{W}_{glove})^{\mathbf{T}}$: Shared weights with input embedding.

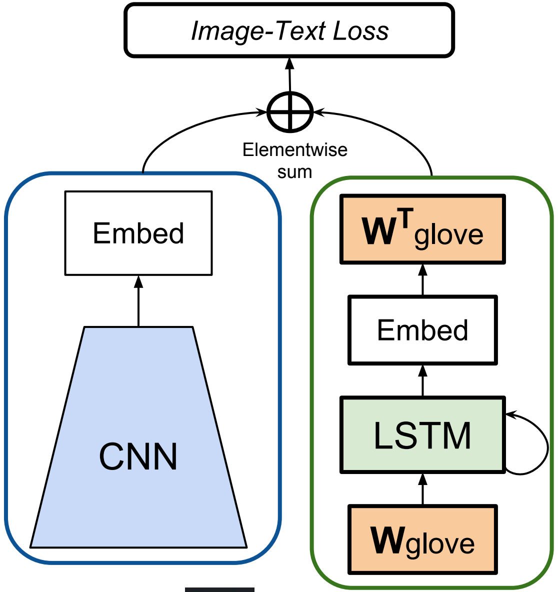
Training Data: Unannotated text data (BNC, ukWac, Wikipedia, Gigaword)

Output: Vector with activations corresponding to scores for *words in the vocabulary*.



Caption Network

Network: Combine output of the visual and text networks. (softmax + cross-entropy loss)

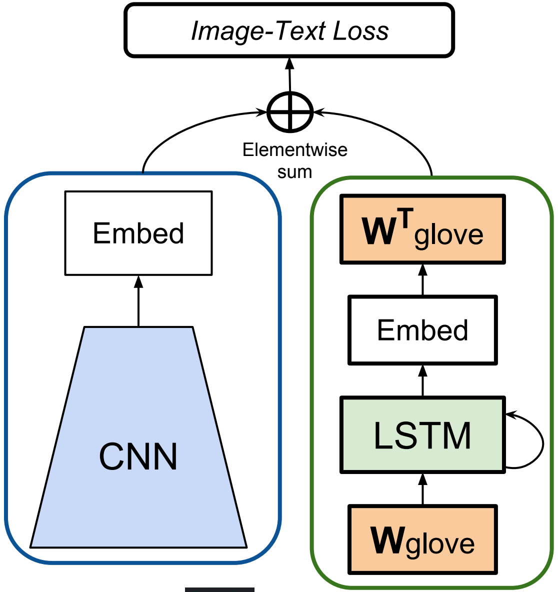


MSCOCO

Caption Model

Training Data:
COCO images with multiple labels

bear, brown, field, grassy, trees, walking

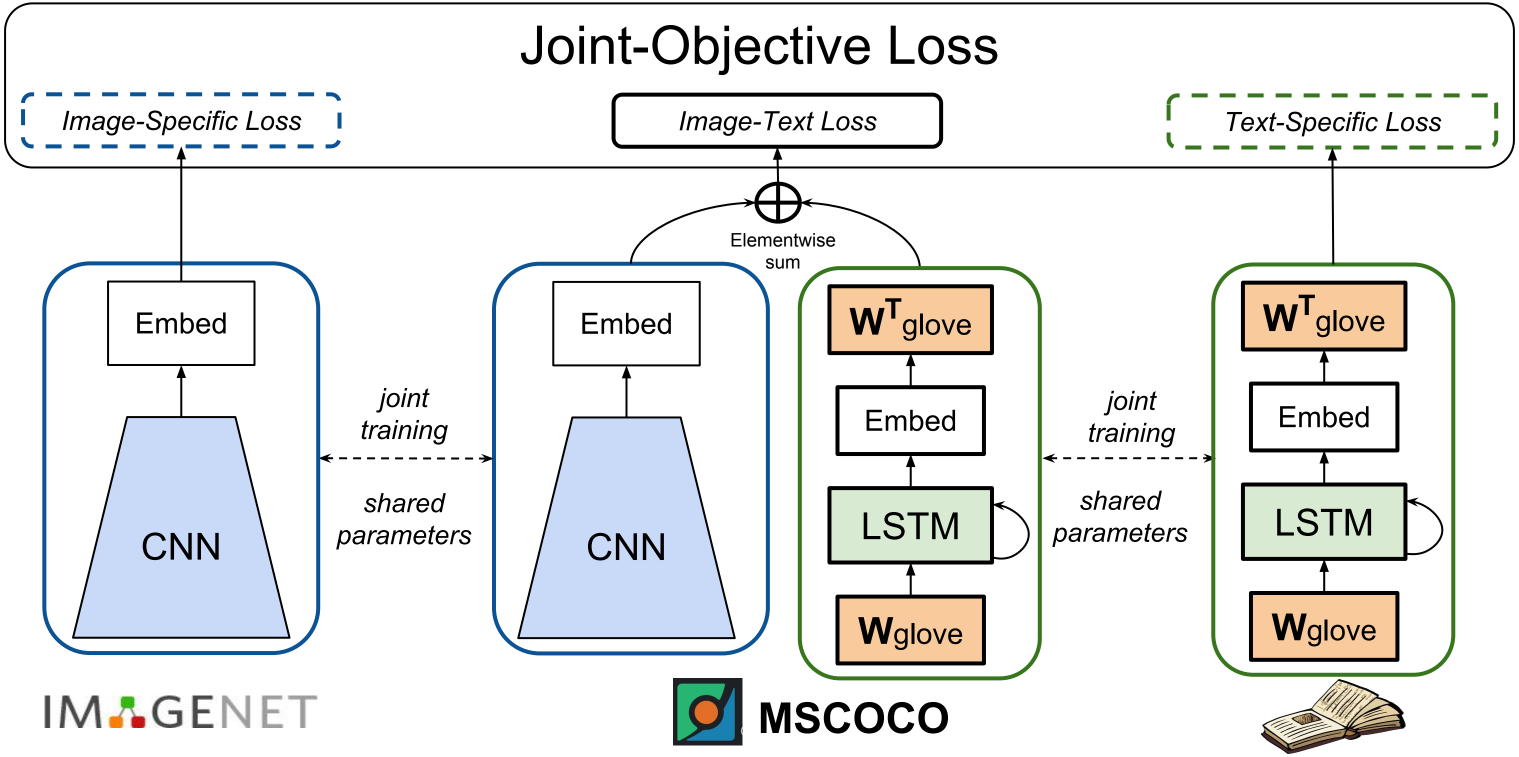


MSCOCO

Training Data:
Captions from COCO

A brown bear walking on a grassy field next to trees

NOC Model: Train simultaneously



Evaluation

- Empirical: COCO held-out objects
  - **In-domain [Use images from COCO]**
  - Out-of-domain [Use imagenet images for held-out concepts]
- Ablations
  - Embedding & Joint training contribution
- ImageNet
  - Quantitative
  - **Human Evaluation - Objects not in COCO**
  - Rare objects in COCO

Evaluation

- Empirical: COCO held-out objects
  - **In-domain [Use images from COCO]**
  - Out-of-domain [Use imagenet images for held-out concepts]
- Ablations
  - Embedding & Joint training contribution
- ImageNet
  - Quantitative
  - **Human Evaluation - Objects not in COCO**
  - Rare objects in COCO

Empirical Evaluation: COCO dataset In-Domain setting

MSCOCO Unpaired Image Data

Elephant, Galloping, Green, Grass

People, Playing, Ball, Field

Black, Train, Tracks

Eat, Pizza

Kitchen, Microwave

MSCOCO Paired Image-Sentence Data

”An elephant galloping in the green grass”

”Two people playing ball in a field”

”A black train stopped on the tracks”

”Someone is about to eat some pizza”

”A kitchen counter with a microwave on it”

MSCOCO Unpaired Text Data

”An elephant galloping in the green grass”

”Two people playing ball in a field”

”A black train stopped on the tracks”

”Someone is about to eat some pizza”

”A microwave is sitting on top of a kitchen counter ”

Empirical Evaluation: COCO heldout dataset

MSCOCO Unpaired Image Data

Elephant, Galloping, Green, Grass

People, Playing, Ball, Field

Black, Train, Tracks

Pizza

Microwave

MSCOCO Paired Image-Sentence Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

"Someone is about to eat some pizza"

"A kitchen counter with a microwave on it"

Held-out

MSCOCO Unpaired Text Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

"A white plate topped with cheesy pizza and toppings."

"A white refrigerator, stove, oven dishwasher and microwave"

Empirical Evaluation: COCO

MSCOCO Unpaired Image Data

Two, elephants, Path, walking

Baseball, batting, boy, swinging

Black, Train, Tracks

Pizza

Microwave

MSCOCO Paired Image-Sentence Data

"An elephant galloping in the green grass"

"Two people playing ball in a field"

"A black train stopped on the tracks"

MSCOCO Unpaired Text Data

"A small elephant standing on top of a dirt field"

"A hitter swinging his bat to hit the ball"

"A black train stopped on the tracks"

"A white plate topped with cheesy pizza and toppings."

"A white refrigerator, stove, oven dishwasher and microwave"

- CNN is pre-trained on ImageNet

Empirical Evaluation: Metrics

F1 (Utility)**:** Ability to recognize and incorporate new words.
(Is the word/object mentioned in the caption?)

METEOR: Fluency and sentence quality.

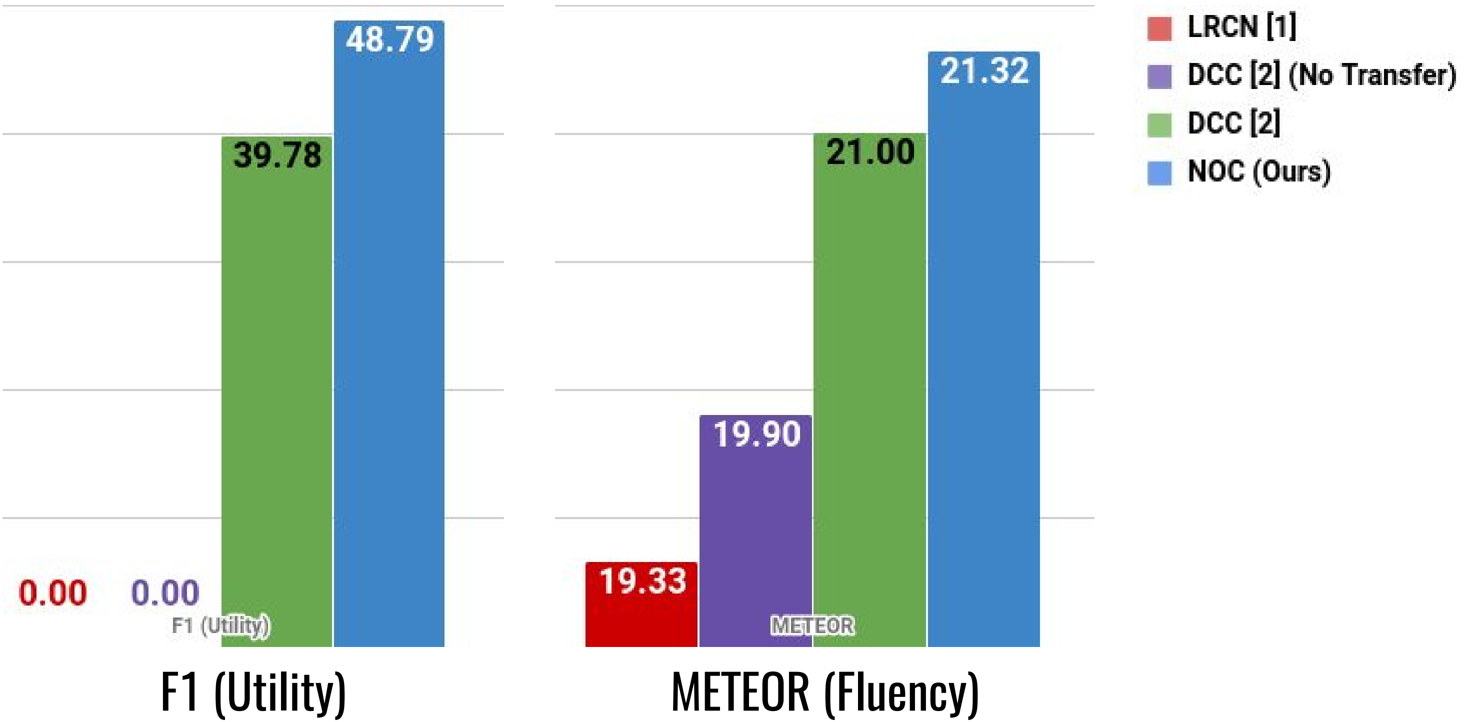
Empirical Evaluation: Baselines

LRCN [1]: Does not caption novel objects.

DCC [2] : Copies parameters for the novel object from a similar object seen in training. (also not end-to-end)

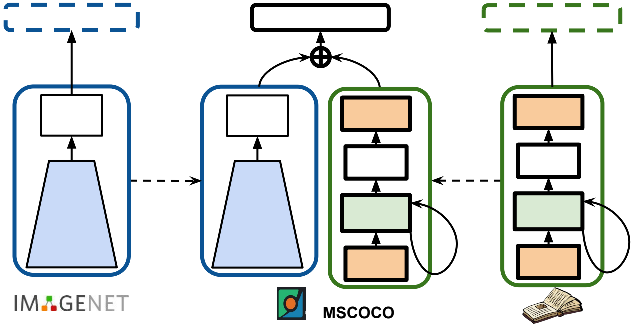
[1] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. CVPR’15
[2] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell CVPR’16

Empirical Evaluation: Results



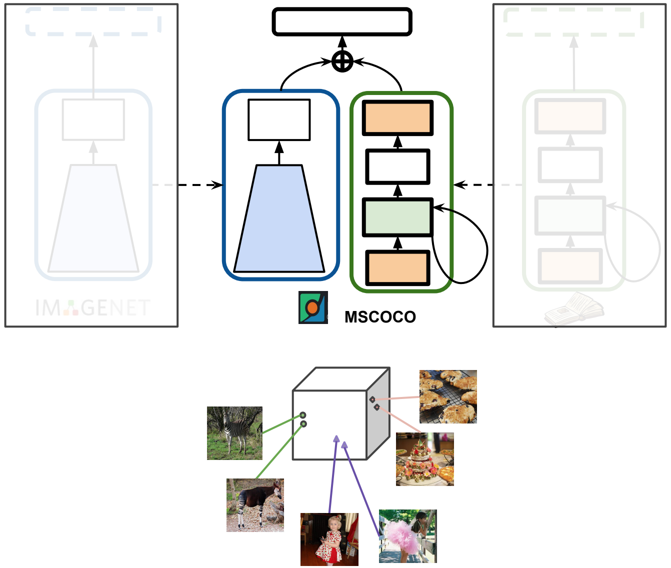
[1] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. CVPR'15
[2] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, T. Darrell CVPR'16

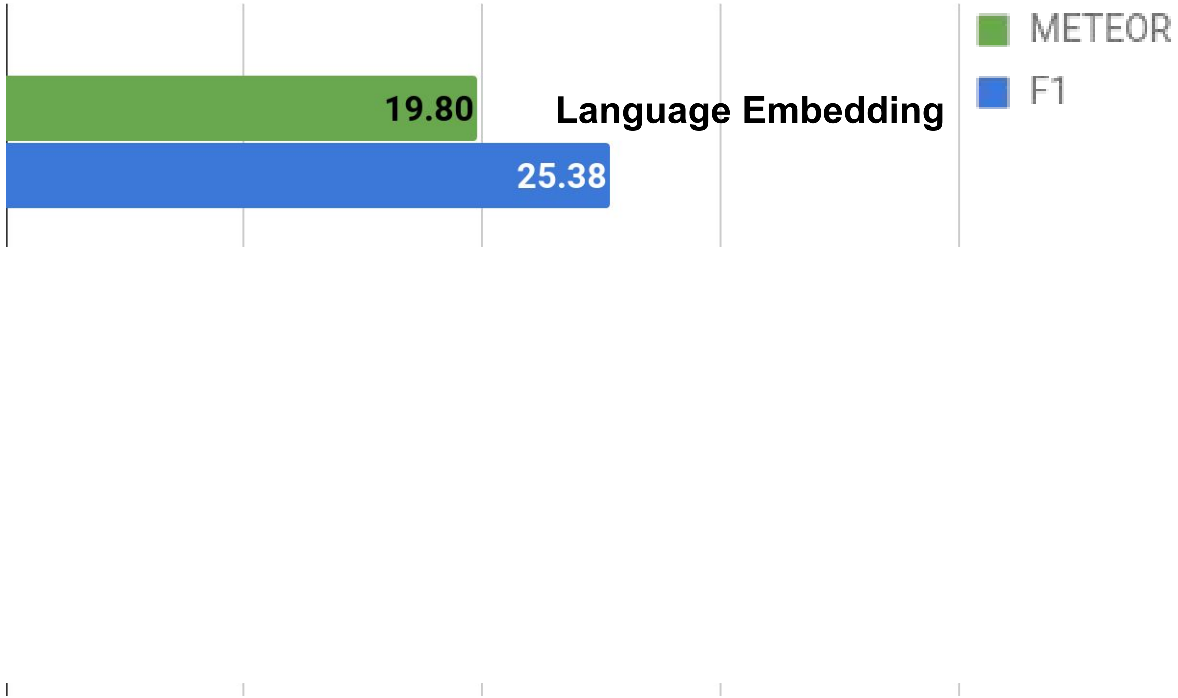
Ablations



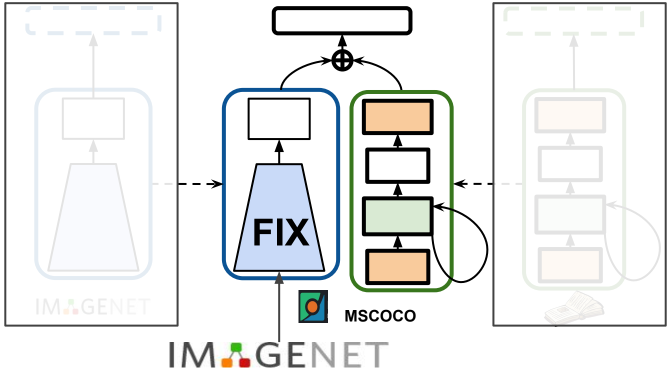
Evaluated on held-out COCO objects.

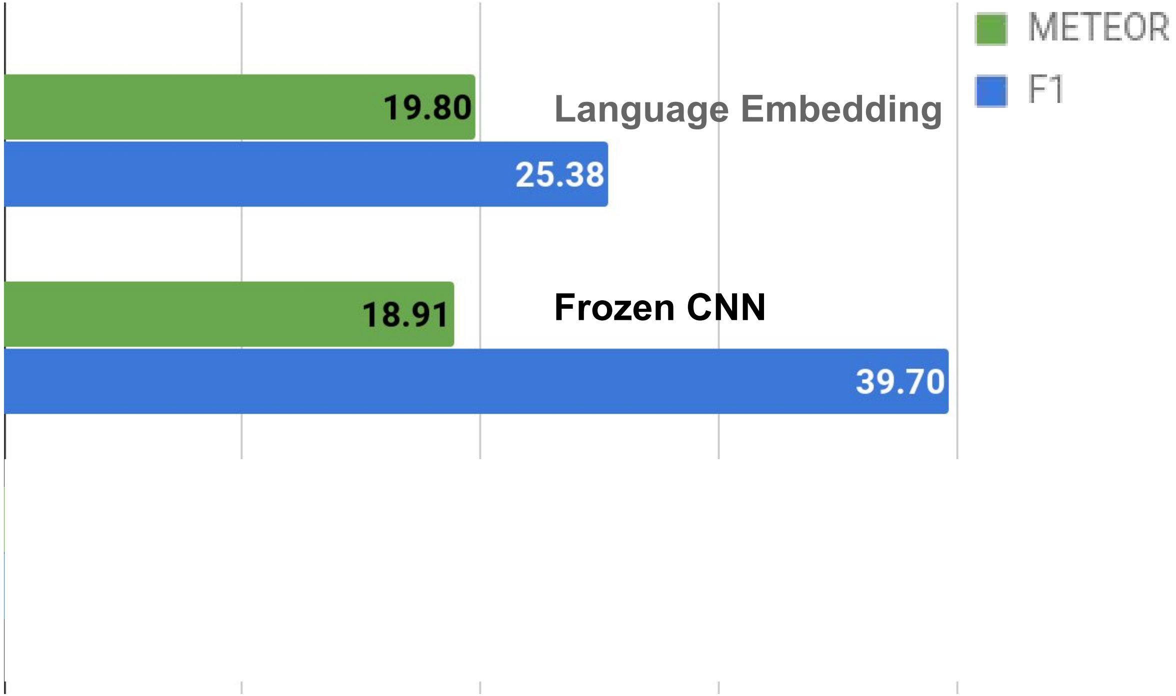
Ablation: Language Embedding





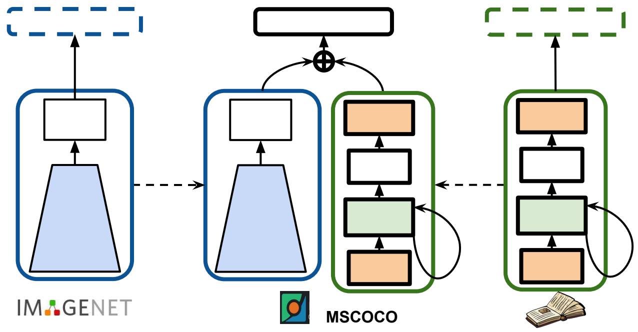
Ablation: Freeze CNN after pre-training

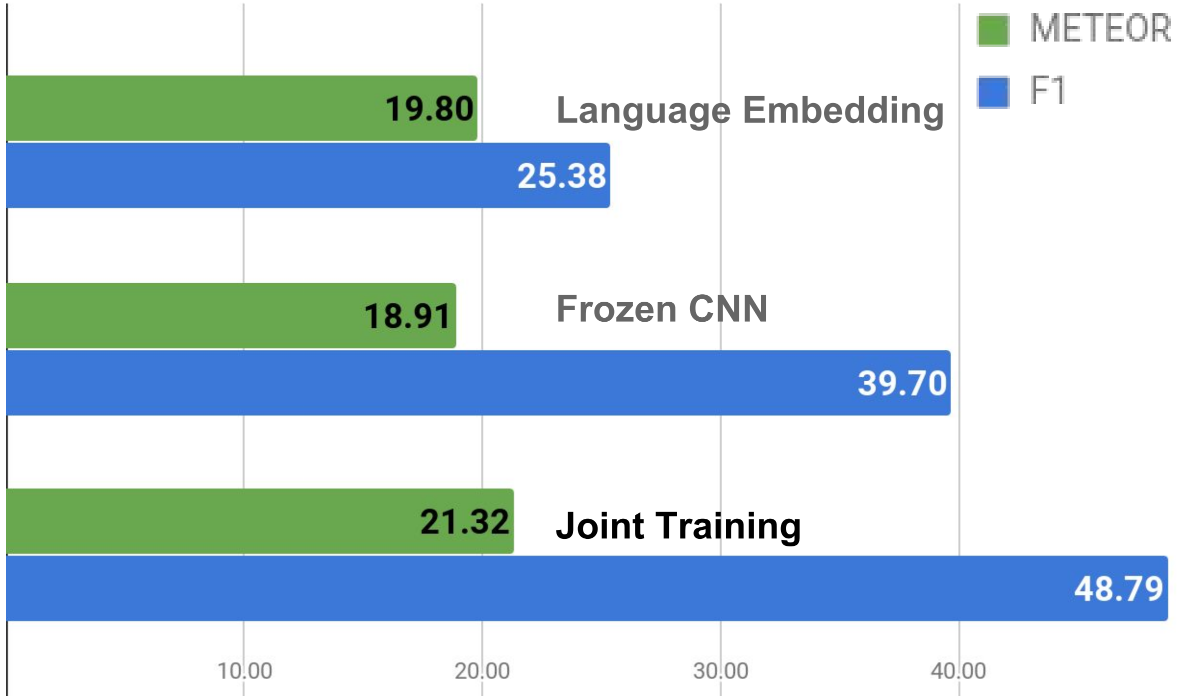




[Catastrophic forgetting in Neural Networks Kirkpatrick et al. PNAS 2017]

Ablation: Joint Training





ImageNet: Human Evaluations

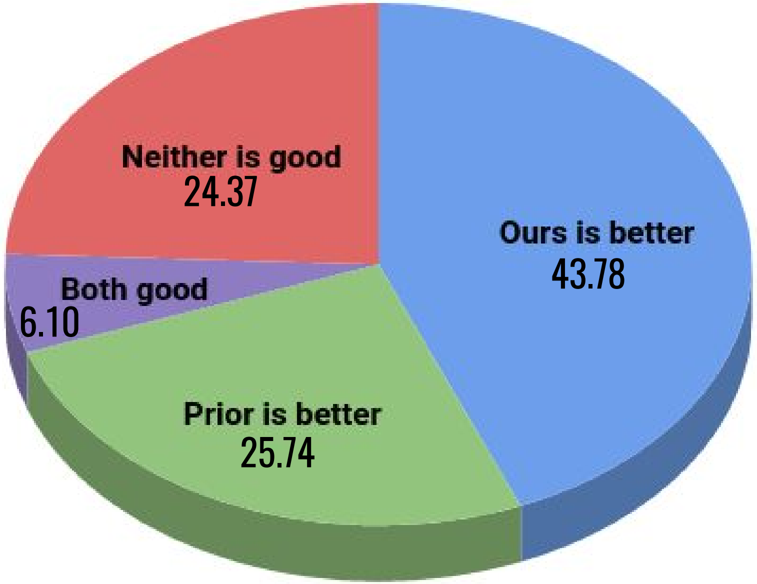
- **ImageNet:** 638 object classes not mentioned in COCO

NOC can describe 582 object classes
(60% more objects than prior work)

ImageNet: Human Evaluations

- **ImageNet:** 638 object classes not mentioned in COCO
- **Word Incorporation:** Which model incorporates the word (name of the object) in the sentence better?
- **Image Description:** Which sentence (model) describes the image better?

ImageNet: Human Evaluations



Word Incorporation

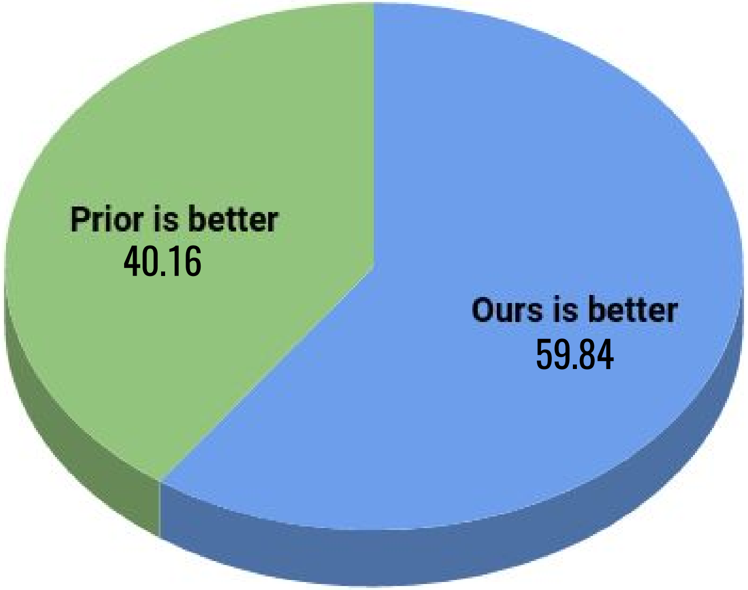
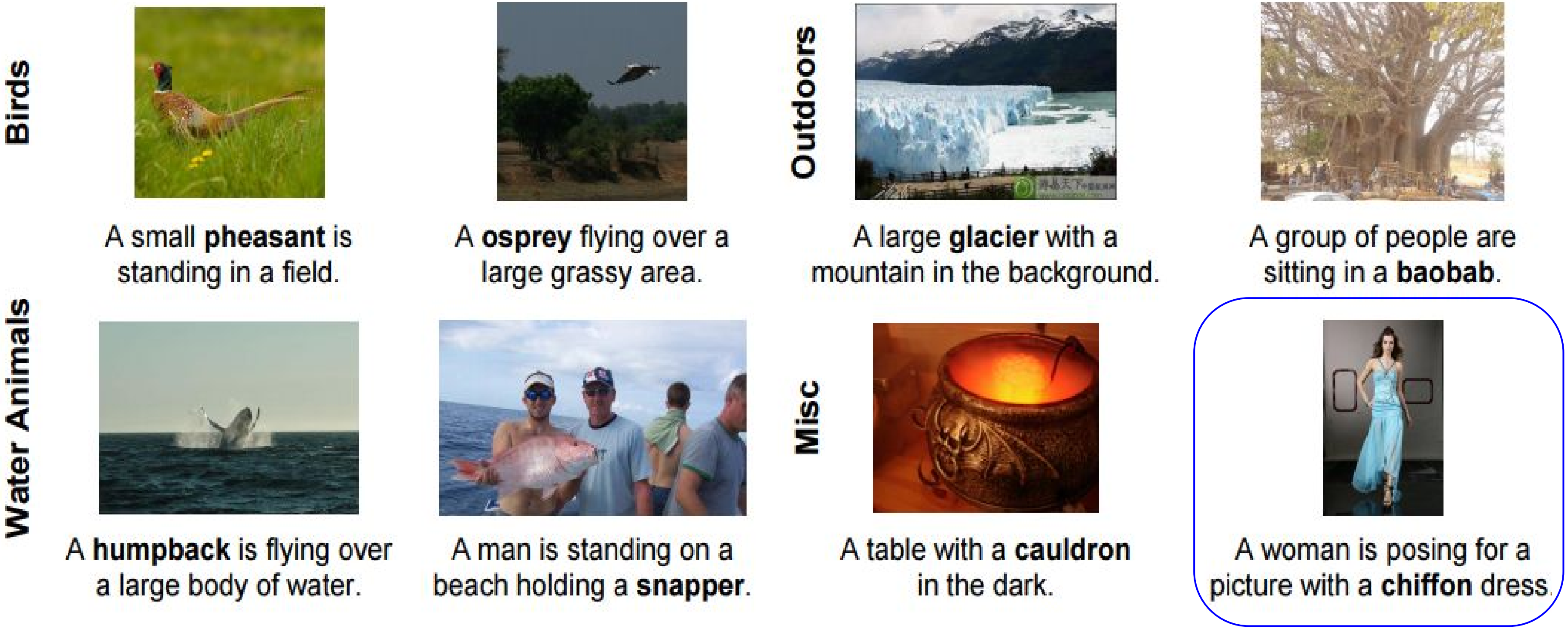


Image Description

Qualitative Evaluation: ImageNet

Qualitative Evaluation: ImageNet



Qualitative Examples: Errors

Balaclava (n02776825) Error: Repetition

NOC: A **balaclava** black and white photo of a man in a **balaclava**.

Sunglass (n04355933) Error: Grammar

NOC: A **sunglass** mirror reflection of a mirror in a mirror.

Gymnast (n10153594) Error: Gender, Hallucination

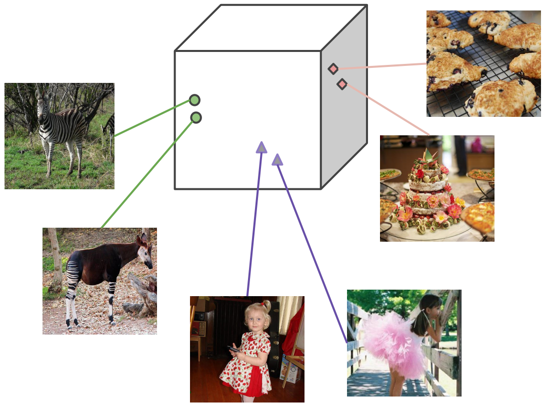
NOC: A man **gymnast** in a blue shirt doing a trick on a skateboard.

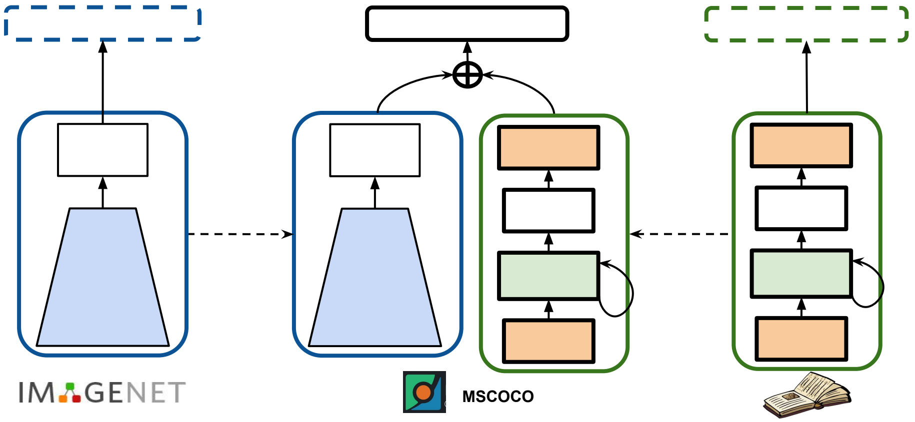
Cougar (n02125311) Error: Description

NOC: A **cougar** with a **cougar** in its mouth.

Novel Object Captioner - Take away

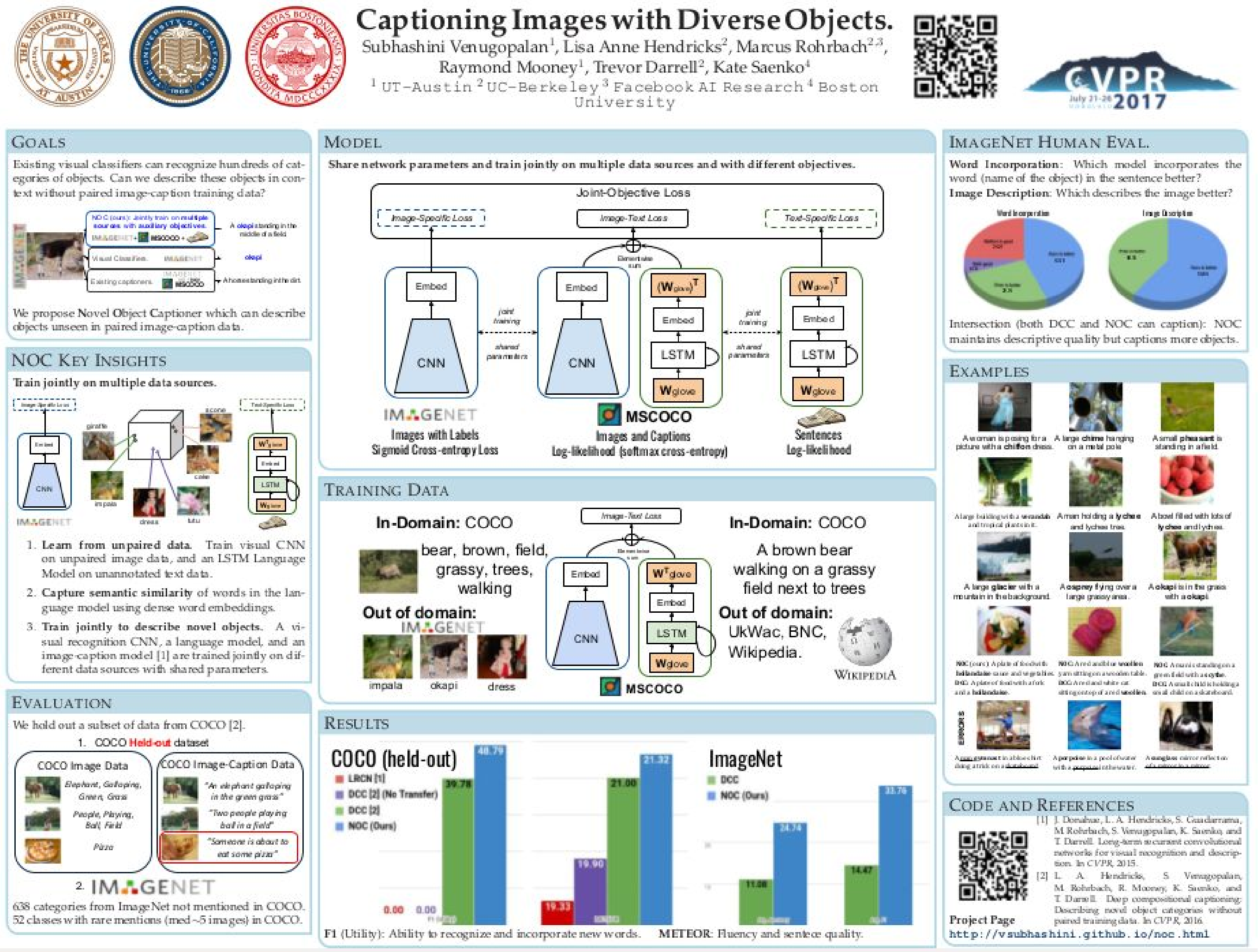
Semantic embeddings and joint training to caption 100s of objects.





A **okapi** standing in the middle of a field.

Poster 11



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
CVPR July 21-26 2017
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 Novel Object Captioner 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
COCO Image Data
COCO Image-Caption Data
"An elephant galloping in the green grass"
"Two people playing ball in a field"
"Someone is about to eat some pizza"
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.
Joint-Objective Loss
Image-Specific Loss
Image-Text Loss
Text-Specific Loss
Embed
CNN
LSTM
joint training
shared parameters
IMAGENET
Images with Labels
Sigmoid Cross-entropy Loss
MSCOCO
Images and Captions
Log-likelihood (softmax cross-entropy)
Sentences
Log-likelihood
TRAINING DATA
In-Domain: COCO
bear, brown, field, grassy, trees, walking
Out of domain:
impala
okapi
dress
In-Domain: COCO
A brown bear walking on a grassy field next to trees
Out of domain:
UkWac, BNC, Wikipedia.
WIKIPEDIA
RESULTS
COCO (held-out)
LRCN [1]
DCC [2] (No Transfer)
DCC [2]
NOC (Ours)
ImageNet
DCC
NOC (Ours)
F1 (Utility): Ability to recognize and incorporate new words.
METEOR: Fluency and sentence 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
[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.
Project Page
http://vsubhashini.github.io/noc.html