Compositional Captioning: Describing Novel Object Categories without Paired Training Data

MLSLP 2016

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

[1] University of California, Berkeley

[2] University of Texas at Austin

[3] Boston University

Visual Description

Berkeley LRCN:
A brown bear standing on top of a lush green field.

MS CaptionBot:
A large brown bear walking through a forest.

LRCN: Donahue, Jeff et al. CVPR 2015.
Microsoft CaptionBot: http://captionbot.ai/

A brown bear walking across a lush green field.

A large brown bear walking through a forest.

A brown bear walks in the grass in front of trees.

A brown bear sitting on top of a green field.

A brown bear walking on a grassy field next to trees.

A large brown bear walking across a lush green field.

Problems with Visual Description

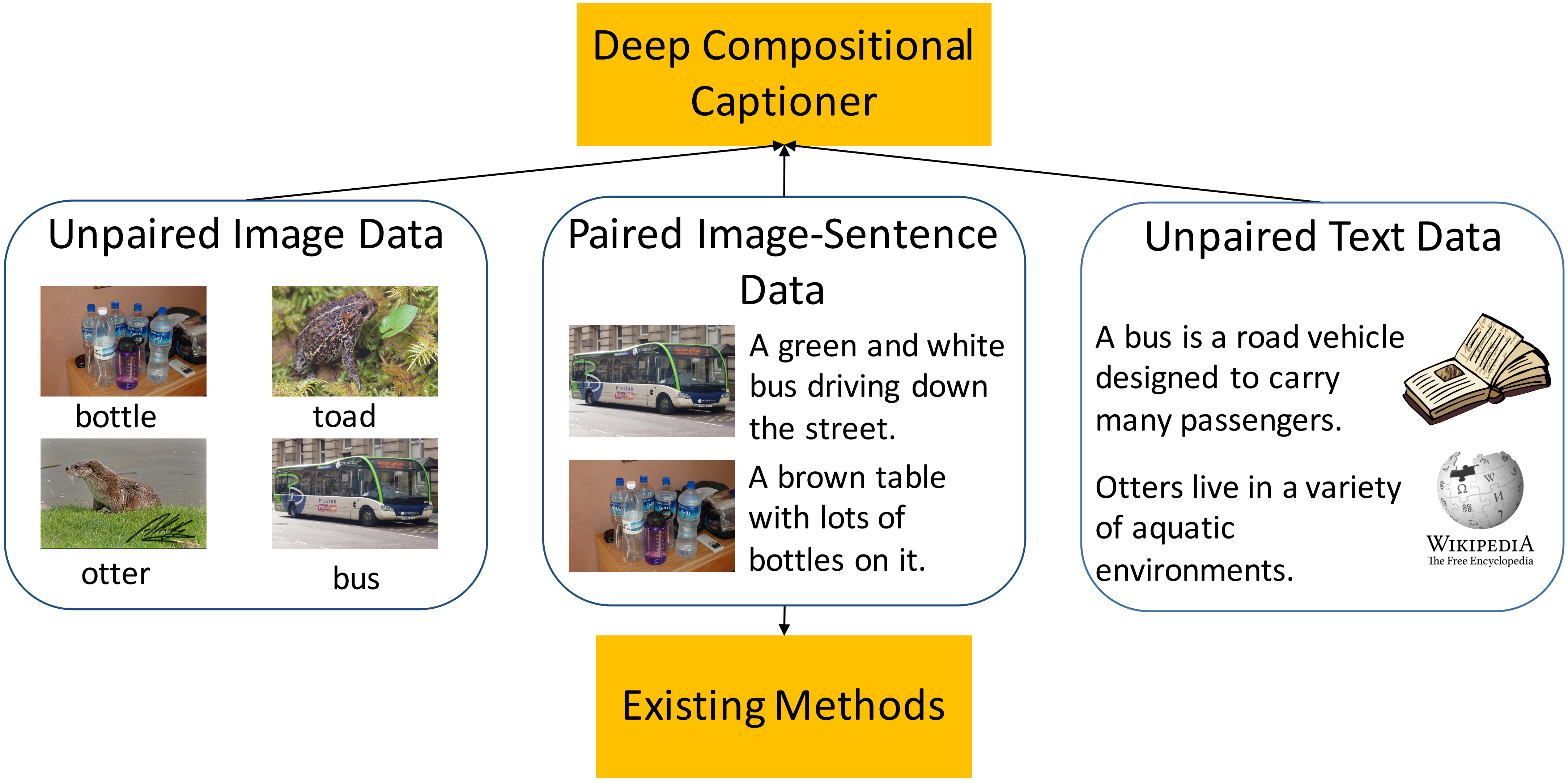
Berkeley LRCN:
"A black bear is standing in the grass."

MS CaptionBot:
"A bear that is eating some grass."

Ours:
"A anteater is standing in the grass."

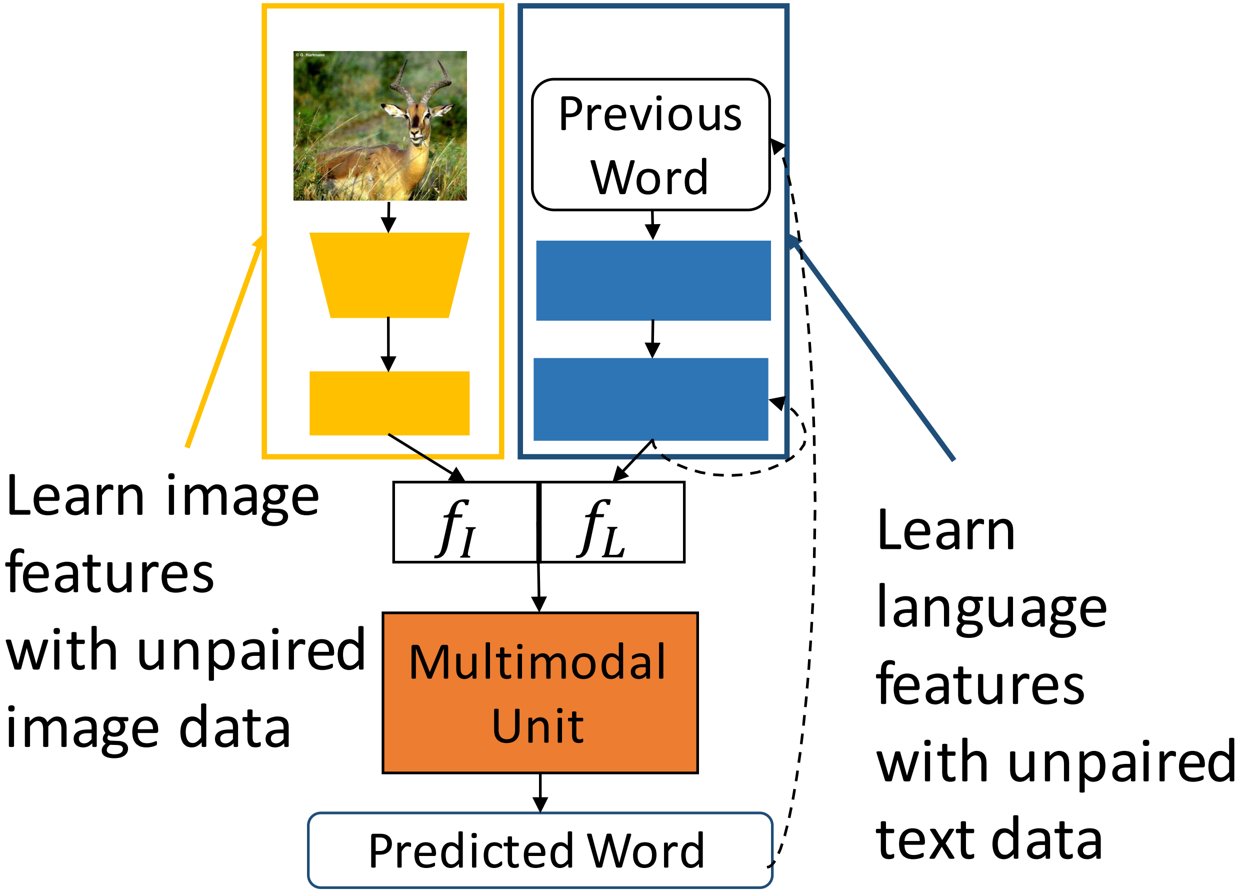
LRCN: Donahue, Jeff et al. CVPR 2015.
CaptionBot: http://captionbot.ai/

We present the **D**eep **C**ompositional **C**aptioner (DCC) which can compose descriptions about novel objects in context.

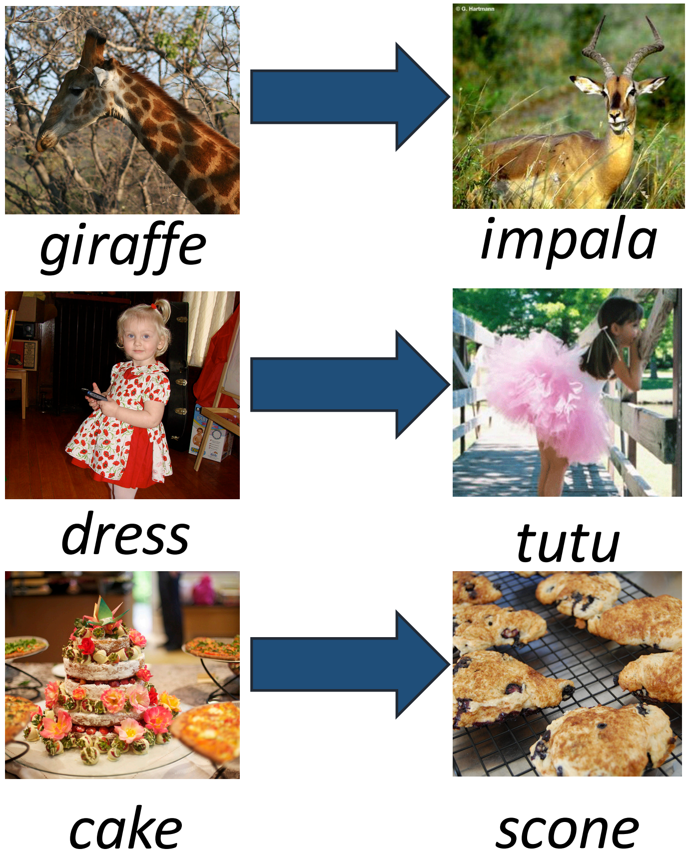


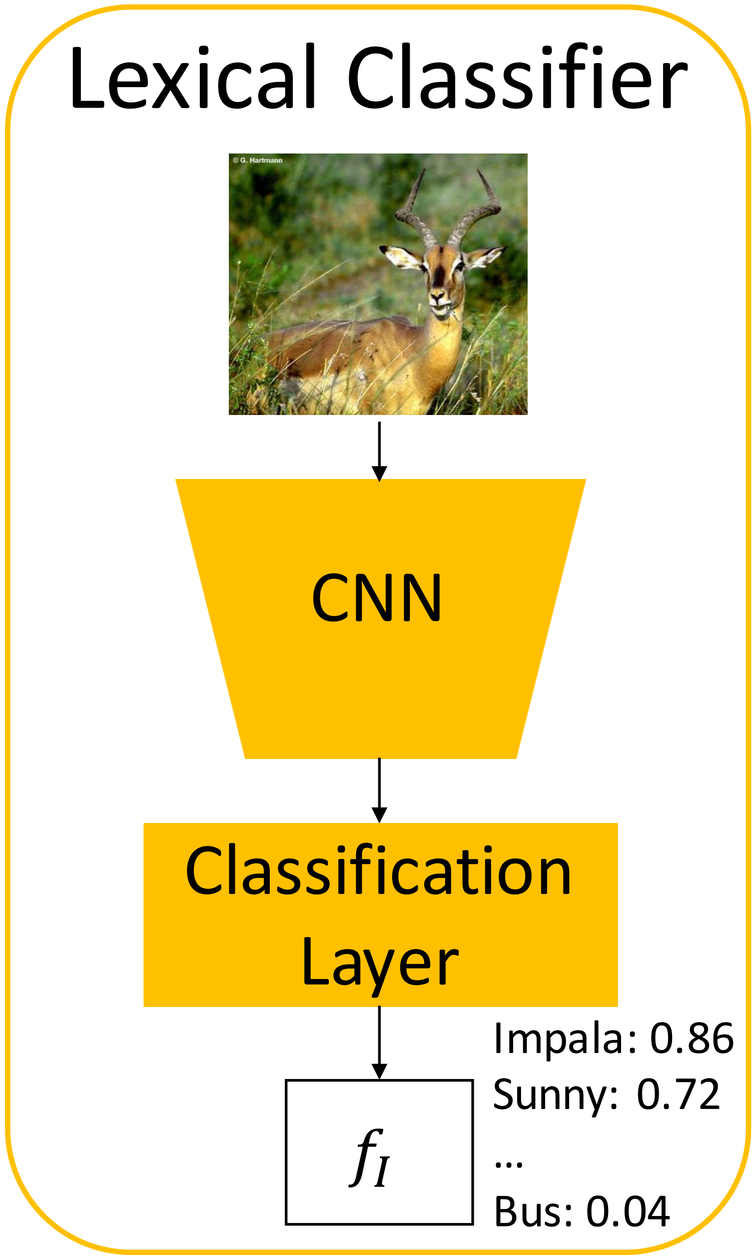
DCC Key Insights

1. Effectively train with outside data



2. Transfer knowledge between related concepts





Training Data: Unpaired Image Data

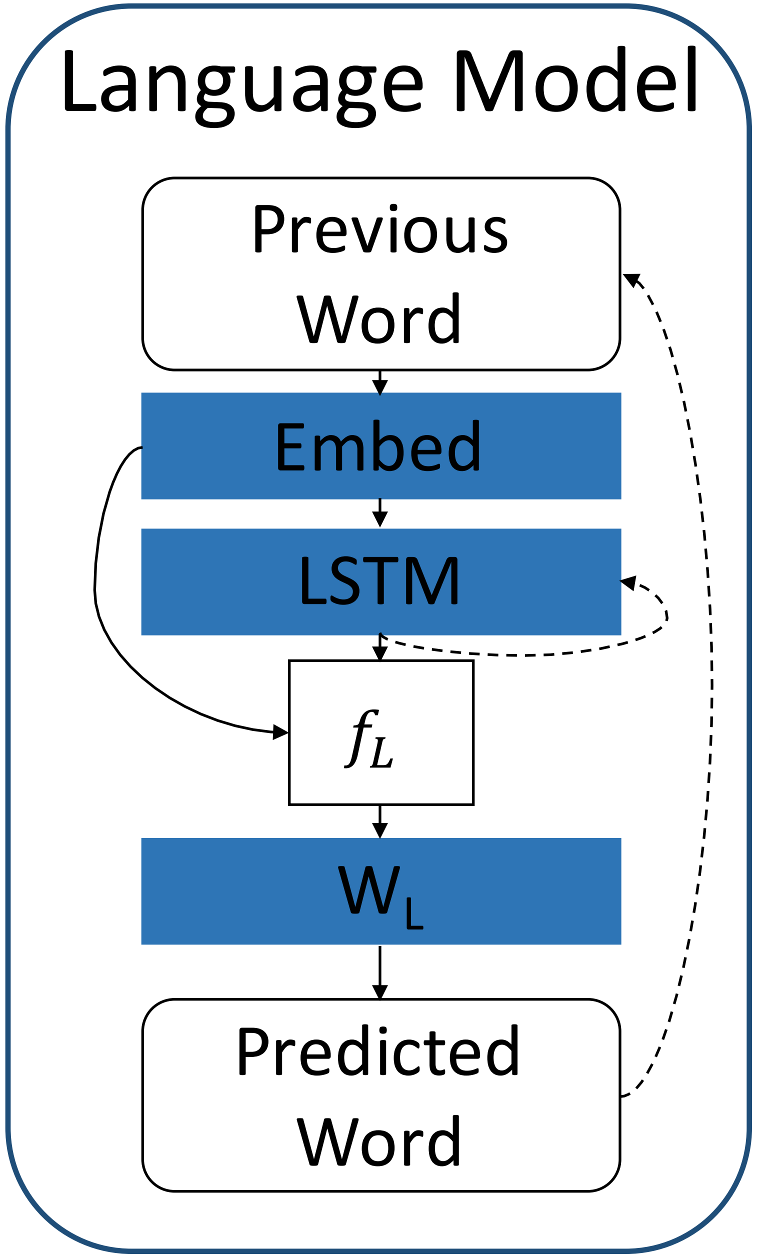
Network: VGG + multilabel loss (sigmoid cross entropy)

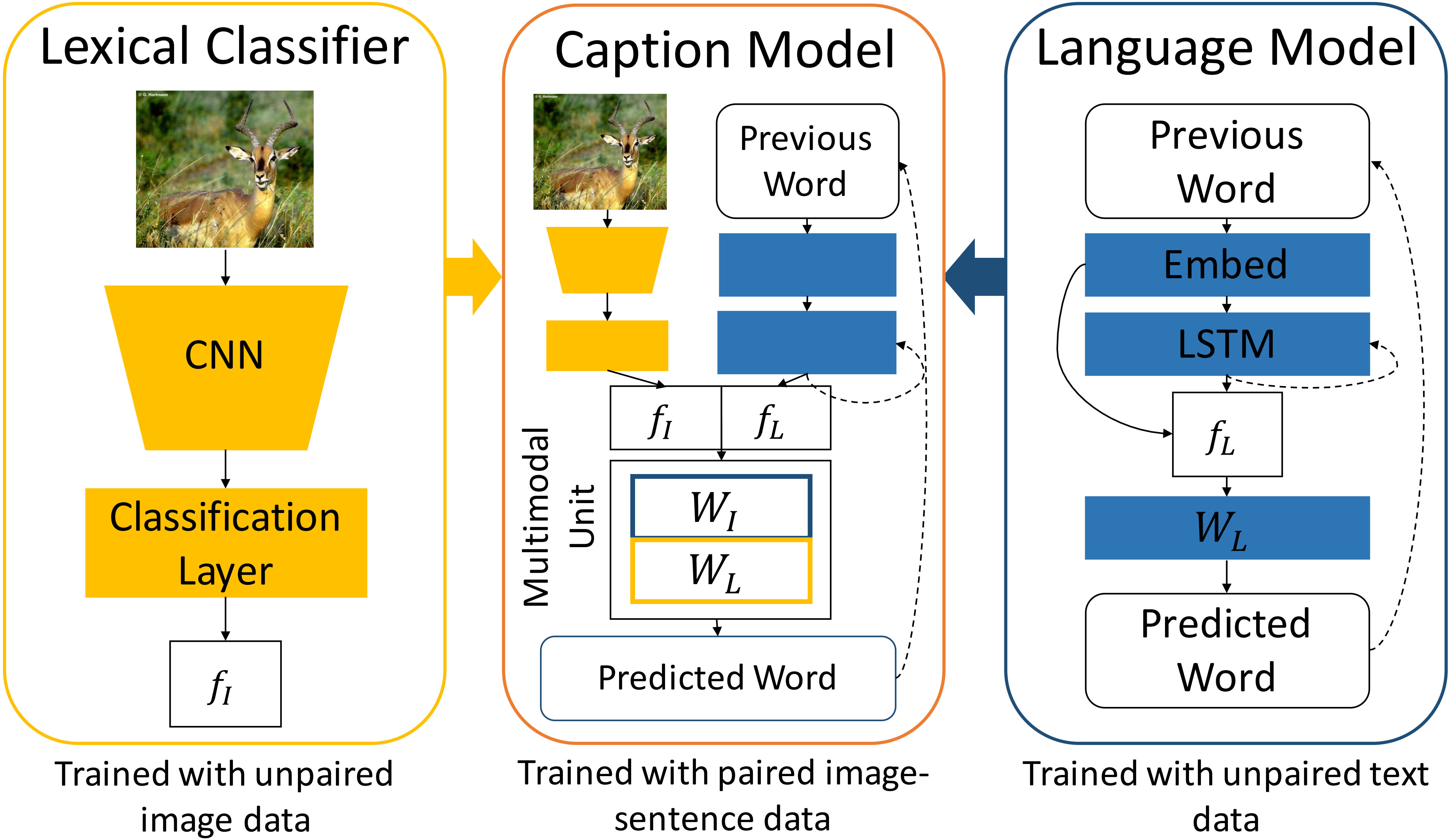
Feature: Vector with activations corresponding to scores for visual concepts in an image.

Training Data: Unpaired Text Data

Network: Embed layer + LSTM unit. Model trained to predict a word, w_t, given the previous words in a sentence, $w_{0:t-1}$.

Feature: Vector which encodes previous words in the sentence.




Lexical Classifier
CNN
Classification Layer
f_I
Caption Model
Previous Word
f_I
f_L
Multimodal Unit
W_I
W_L
Predicted Word
Language Model
Previous Word
Embed
LSTM
f_L
W_L
Predicted Word
Trained with unpaired image data
Trained with paired image-sentence data
Trained with unpaired text data

Multimodal Unit

Language Feature

Image Feature

$$S(w_t | I, w_{0:t}) = \boxed{f_L} W_L + \boxed{f_I} W_I + b$$

$f_L W_L$ large for:

Giraffe

Horse

Couch

…

Standing

$f_I W_I$ large for:

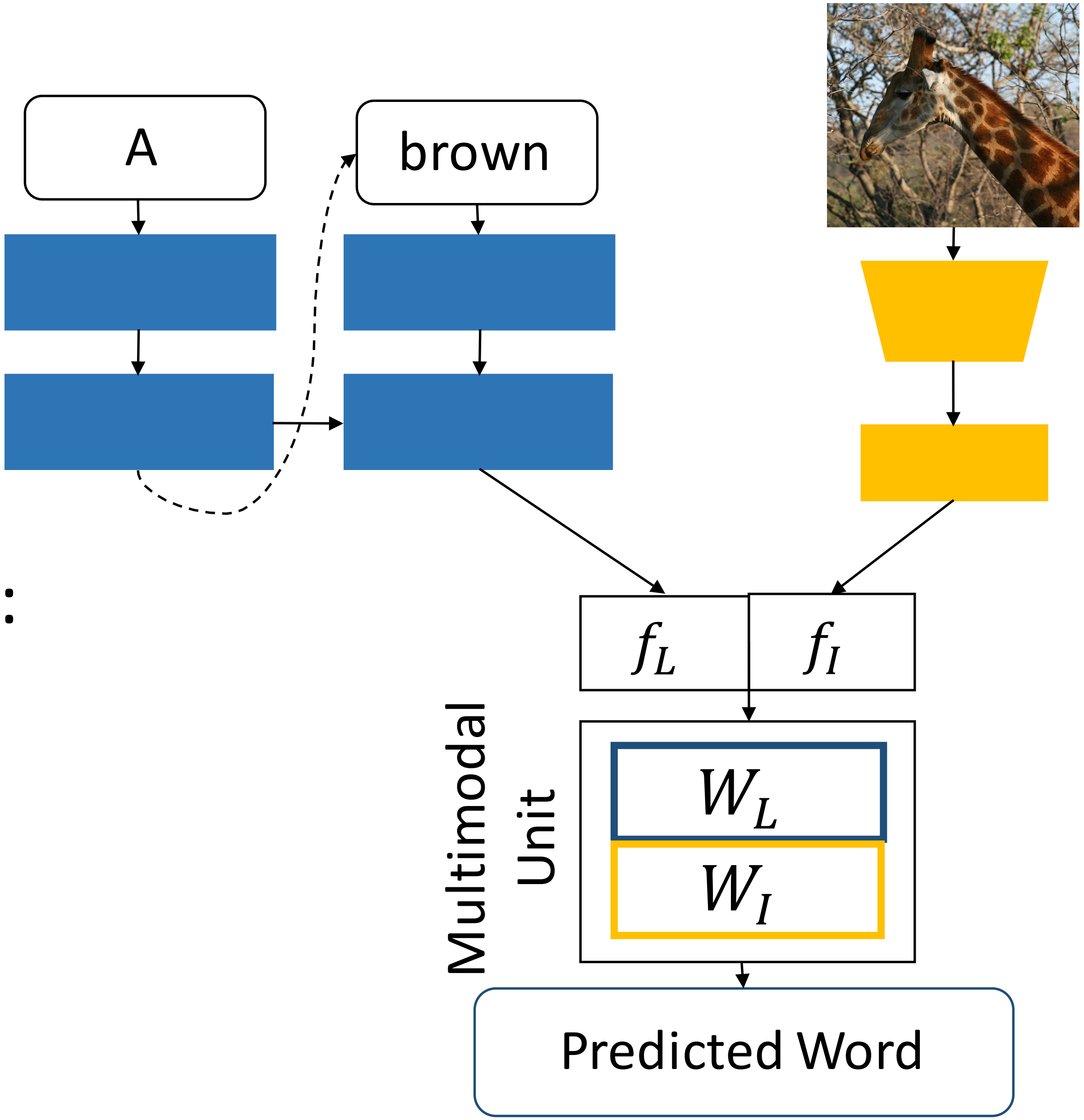
Giraffe

Trees

Standing

…

Couch



Multimodal Unit

$$S(w_t|I, w_{0:t}) = f_L W_L + f_I W_I + b$$

$f_L W_L$ large for:

Giraffe

Horse

Couch

…

Standing

$f_I W_I$ large for:

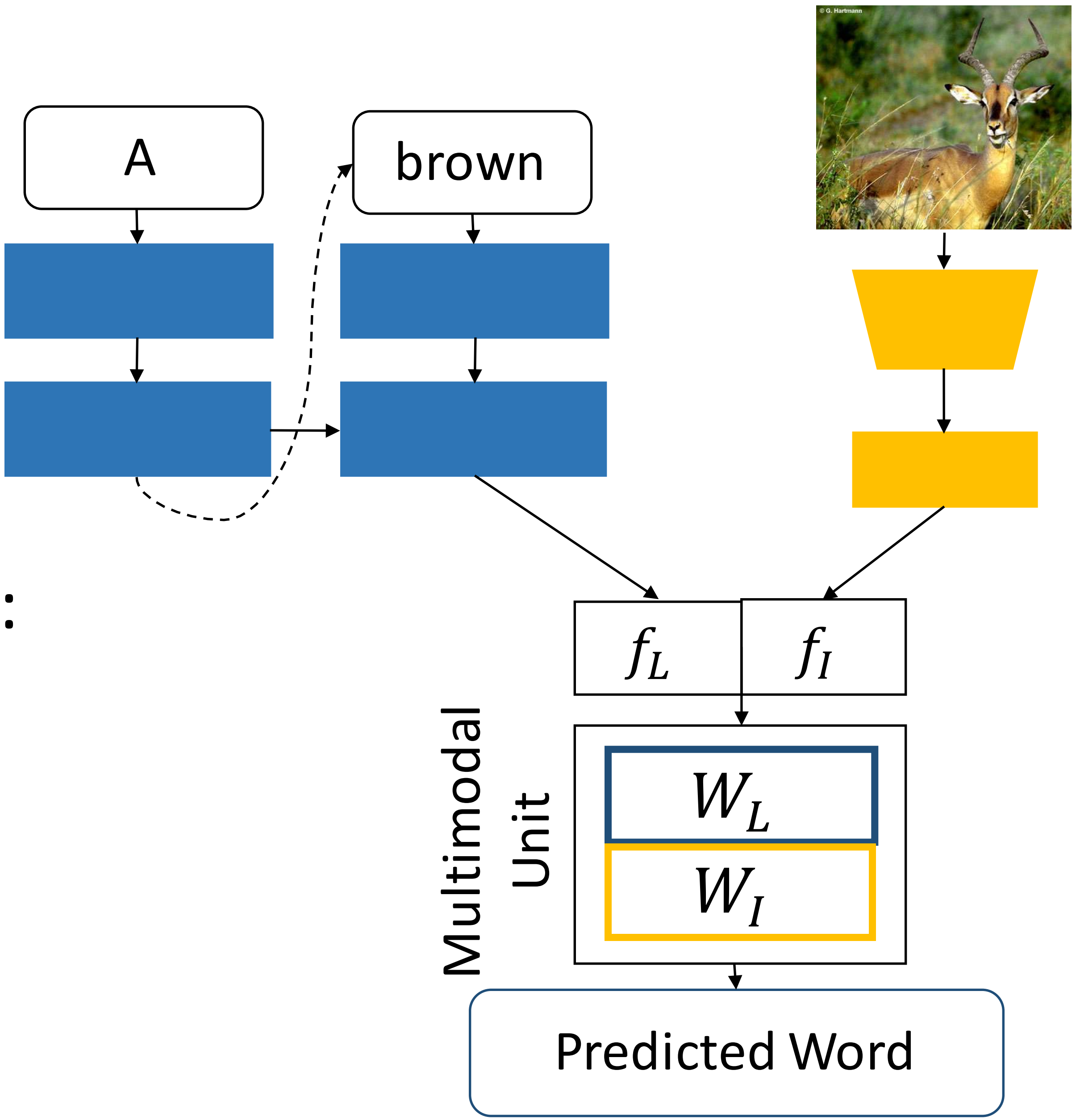
Giraffe

Trees

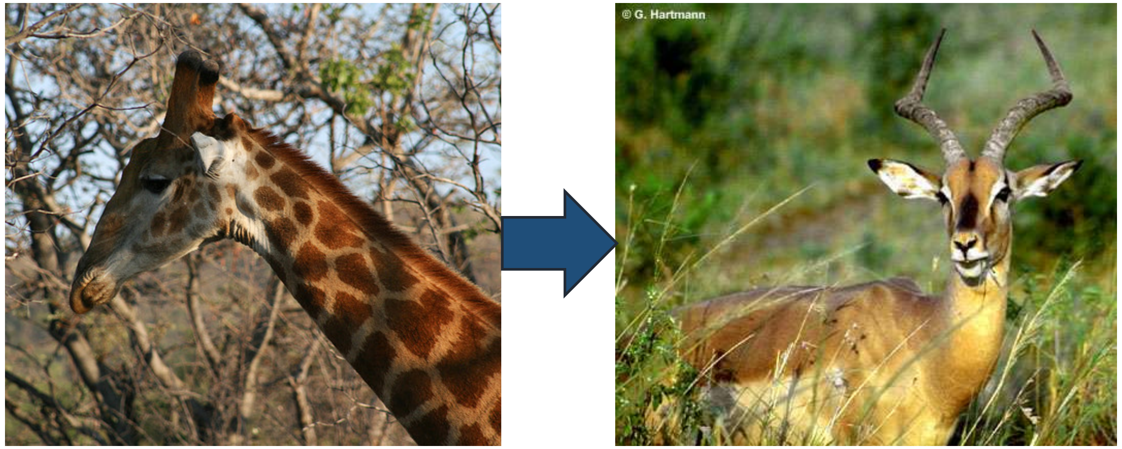
Standing

…

Couch



Weight Transfer

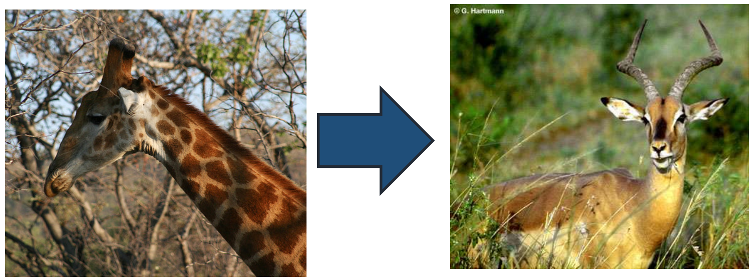


Transfer pair chosen using word2vec

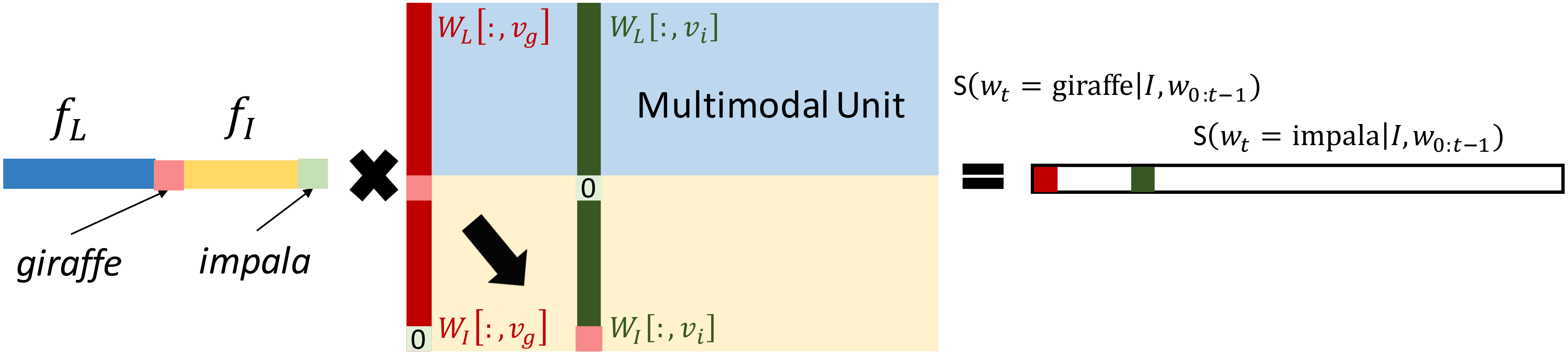
Weight Transfer

$$S(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$

$$S(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec



Evaluation

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 ”

Evaluation

MSCOCO Unpaired Image Data

Elephant, Galloping, Green, Grass

People, Playing, Ball, Field

Black, Train, Tracks

Pizza

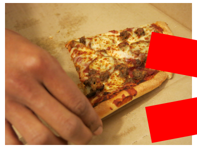
Microwave

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”*~~

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 ”

Held-out dataset

Results: MSCOCO In-Domain

	DCC (Ours)

Comparison of DCC to LRCN and DCC with no transfer.

- High F1 score indicates DCC can describe words outside of paired image sentence data
- Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (Ours)

Comparison of DCC to LRCN and DCC with no transfer.

- ➢ High F1 score indicates DCC can describe words outside of paired image sentence data
- ➢ Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (No Transfer)	DCC (Ours)

Comparison of DCC to LRCN and DCC with no transfer.

- ➢ High F1 score indicates DCC can describe words outside of paired image sentence data
- ➢ Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (No Transfer)	DCC (Ours)
Efficacy (F1)			

Comparison of DCC to LRCN and DCC with no transfer.

- High F1 score indicates DCC can describe words outside of paired image sentence data
- Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (No Transfer)	DCC (Ours)
Efficacy (F1)			
Sentence Quality (METEOR)			

Comparison of DCC to LRCN and DCC with no transfer.

- ➢ High F1 score indicates DCC can describe words outside of paired image sentence data
- ➢ Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (No Transfer)	DCC (Ours)
Efficacy (F1)	0.00	0.00	**39.78**
Sentence Quality (METEOR)			

Comparison of DCC to LRCN and DCC with no transfer.

- High F1 score indicates DCC can describe words outside of paired image sentence data
- Increased METEOR indicates DCC produces better sentences

Results: MSCOCO In-Domain

	LRCN	DCC (No Transfer)	DCC (Ours)
Efficacy (F1)	0.00	0.00	**39.78**
Sentence Quality (METEOR)	19.33	19.90	**21.00**

Comparison of DCC to LRCN and DCC with no transfer.

- High F1 score indicates DCC can describe words outside of paired image sentence data
- Increased METEOR indicates DCC produces better sentences

Empirical Evaluation

MSCOCO Unpaired Image Data

Elephant, Galloping, Green, Grass

People, Playing, Ball, Field

Black, Train, Tracks

IMAGENET

Pizza

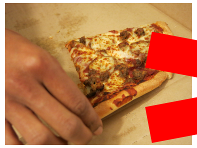
Microwave

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 i[illegible]bout to [illegible] pizza”*~~

~~*”A kitchen co[illegible]ter with [illegible]ve 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”

WIKIPEDIA
The Free Encyclopedia

”Pepperoni is a popular pizza topping.”

”All microwaves use a timer for the cooking time”

Out-of-Domain Held Out Dataset

Results: MSCOCO Out-Of-Domain

	Unpaired Image Data	Unpaired Text Data	METEOR	F1
LRCN	N/A	N/A	19.33	0.00
DCC (No Transfer)	MSCOCO	MSCOCO	19.90	0.00
DCC (Ours)	MSCOCO	MSCOCO	21.00	39.78

DCC performs well when using out of domain data to train the lexical classifier and language model.

Results: MSCOCO Out-Of-Domain

	Unpaired Image Data	Unpaired Text Data	METEOR	F1
LRCN	N/A	N/A	19.33	0.00
DCC (No Transfer)	MSCOCO	MSCOCO	19.90	0.00
DCC (Ours)	MSCOCO	MSCOCO	21.00	39.78
DCC (Ours)	ImageNet	MSCOCO	20.71	33.60

DCC performs well when using out of domain data to train the lexical classifier and language model.

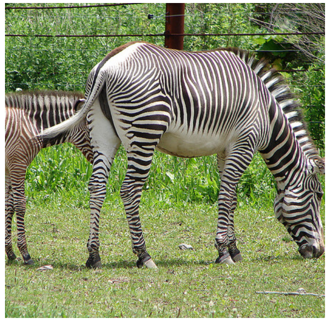
Results: MSCOCO Out-of-Domain

	Unpaired Image Data	Unpaired Text Data	METEOR	F1
LRCN	N/A	N/A	19.33	0.00
DCC (No Transfer)	MSCOCO	MSCOCO	19.90	0.00
DCC (Ours)	MSCOCO	MSCOCO	21.00	39.78
DCC (Ours)	ImageNet	MSCOCO	20.71	33.60
DCC (Ours)	ImageNet	CaptionTxt	20.66	35.53
DCC (Ours)	ImageNet	WebCorpus	20.66	34.94

DCC performs well when using out of domain data to train the lexical classifier and language model.

No transfer: *A green and white street sign on a city street.*
DCC: *A green and white* ***bus*** *parked on the side of the street.*

No transfer: *A dog lying on a bed with a large brown dog.*
DCC: *A dog lying on a* ***couch*** *with a large window in the background.*



No transfer: *Two giraffes are eating grass in the field.*
DCC: *Two* ***zebra*** *grazing in a green grass field.*

No transfer: *A white and black cat is sitting on a toilet.*
DCC: *A white* ***microwave*** *sitting on a brick wall.*

DCC can describe over 300 ImageNet visual concepts in diverse contexts.

DCC:
A person is holding a **gecko** in their hand.

Berkeley LRCN:
A person holding a piece of food in their hand.

MS CaptionBot:
A close up of a person holding a baby.

DCC:
A **gecko** is standing on a branch of a tree.

Berkeley LRCN:
A bird is standing on the edge of a rock.

MS CaptionBot:
A bird that is standing in the water.

DCC can describe over 300 ImageNet visual concepts in diverse contexts.

A woman in a **chiffon tutu**.

A bunch of a **lychee** are in a market.

A black and white photo of a **candelabra** in a room.

A group of people standing around a **baobab** in a field.

A close up of a wooden table with a bottle of **whisky**.

A brown **bobcat** in a green field.

A close up of a **scone** on a plate.

A white **centrifuge** is sitting on the table.

Failure Cases

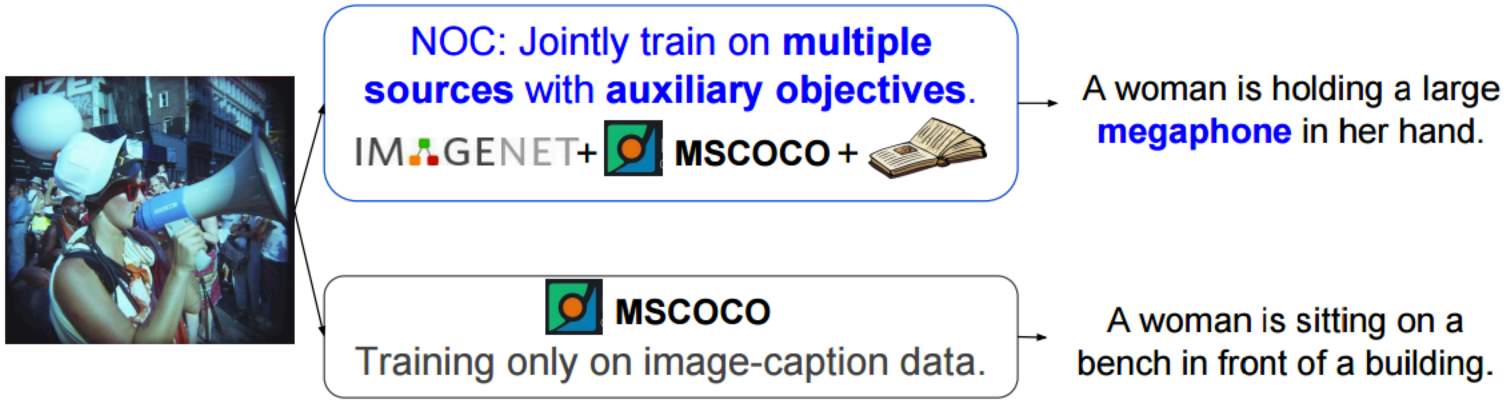
A woman is riding a **unicycle** on a **unicycle**.

A group of people standing around a **foxhunting** on a field.

Results: Video Description

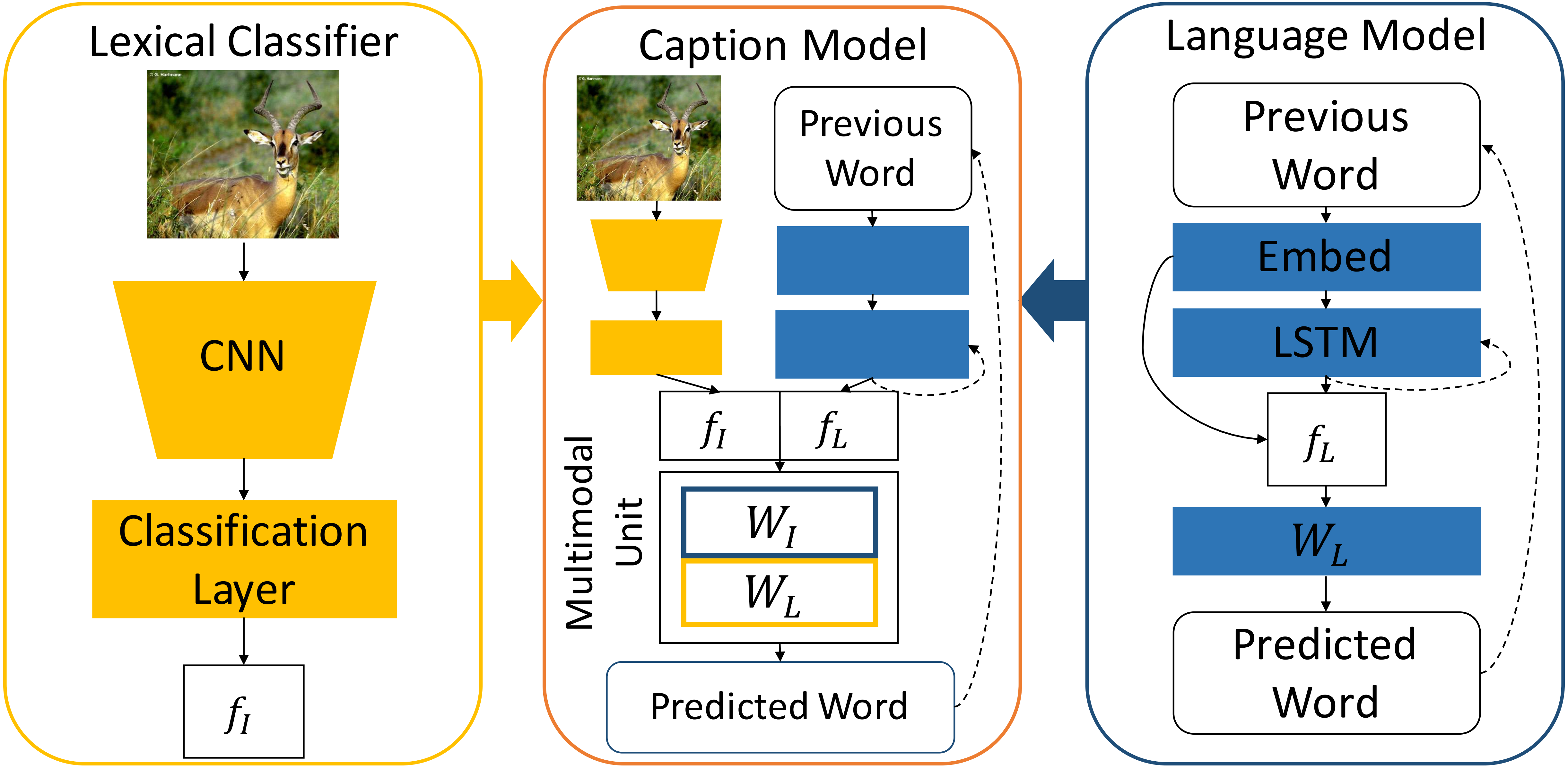
	METEOR	F1
Baseline (No Transfer)	28.80	0.0
+DCC (ours)	28.9	6.0
+ ILSVRC Videos (No Transfer)	29.0	0.0
+ DCC (ours) + ILSVRC Videos	**29.10**	**22.2**

Novel Object Captioner

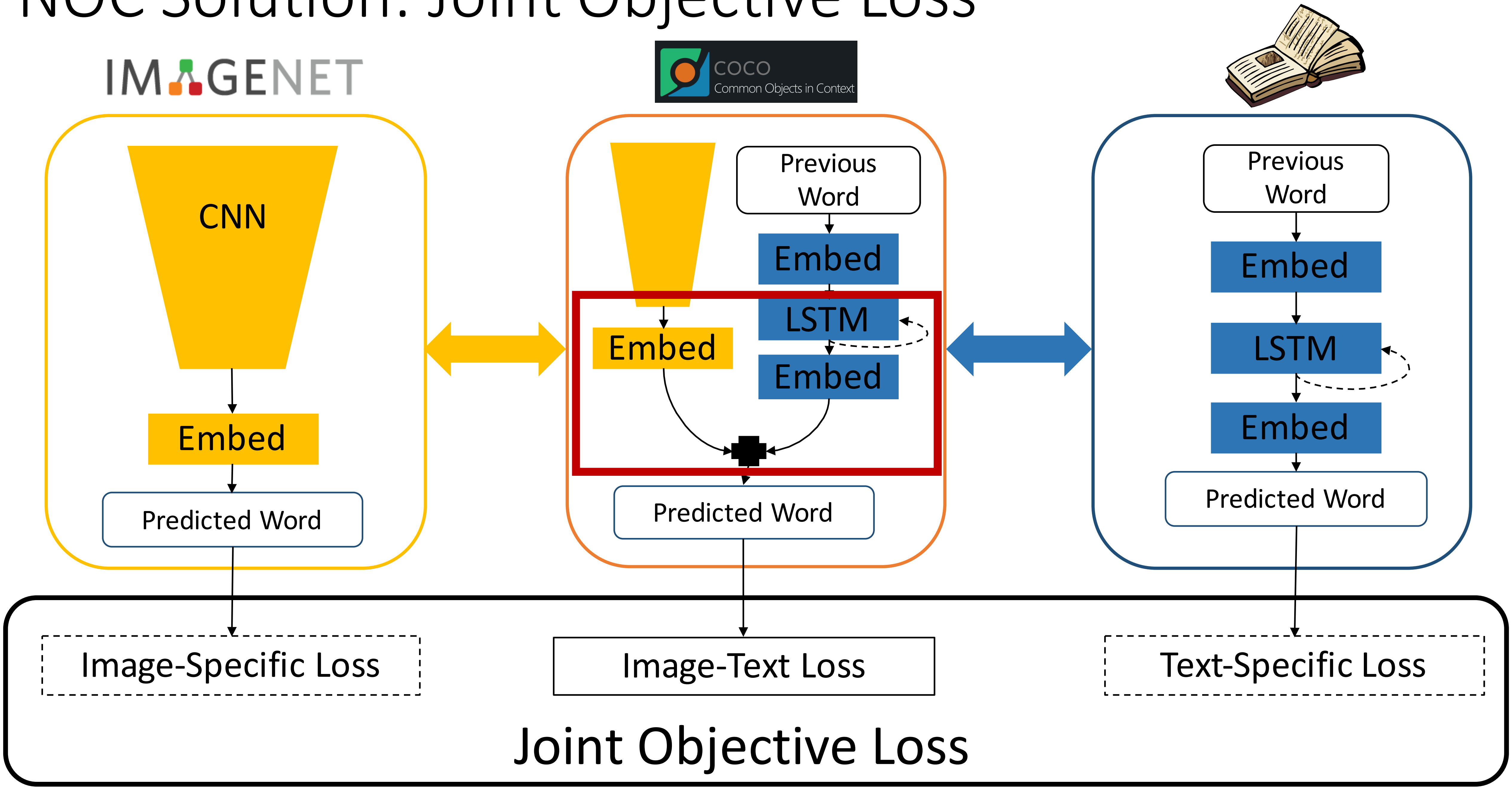


"Captioning Images with Diverse Objects" Venugopalan 2016
http://arxiv.org/abs/1606.07770

DCC Issue: Not End-to-End Trainable



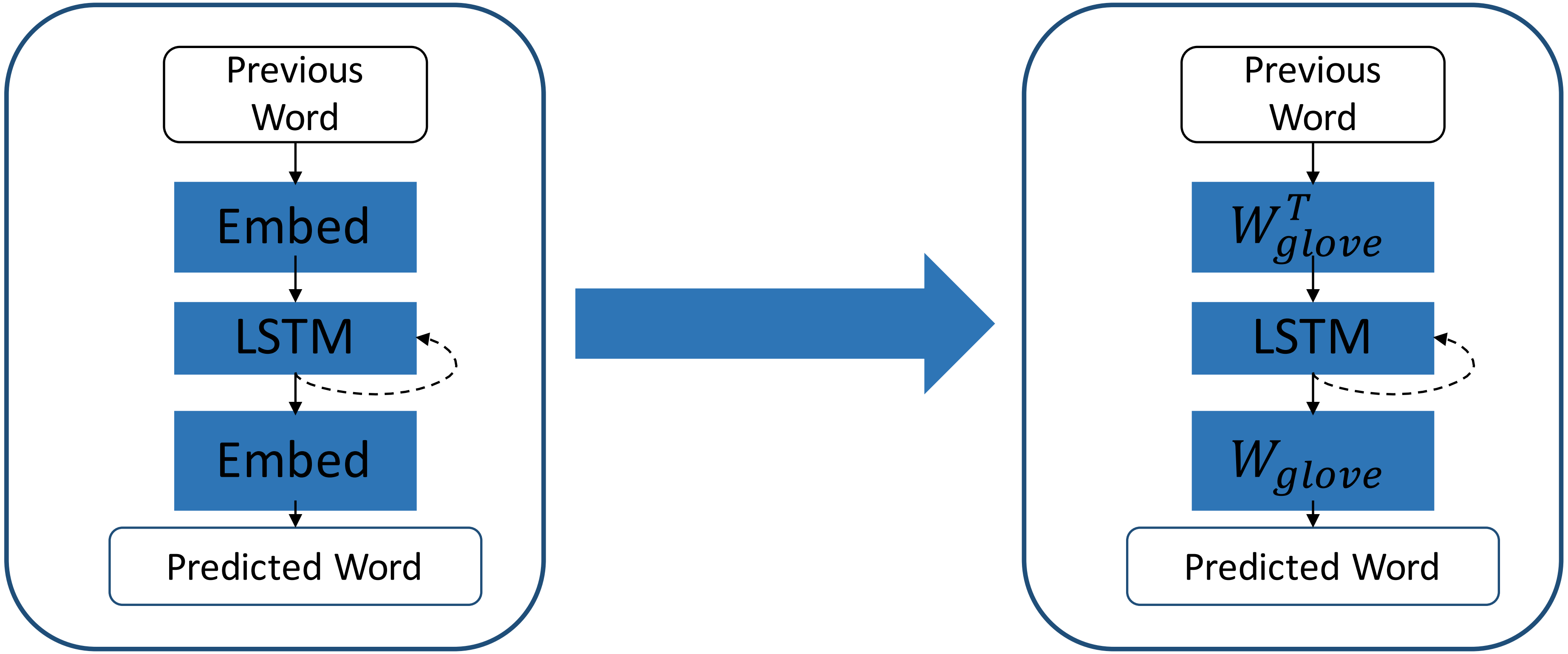
NOC Solution: Joint Objective Loss



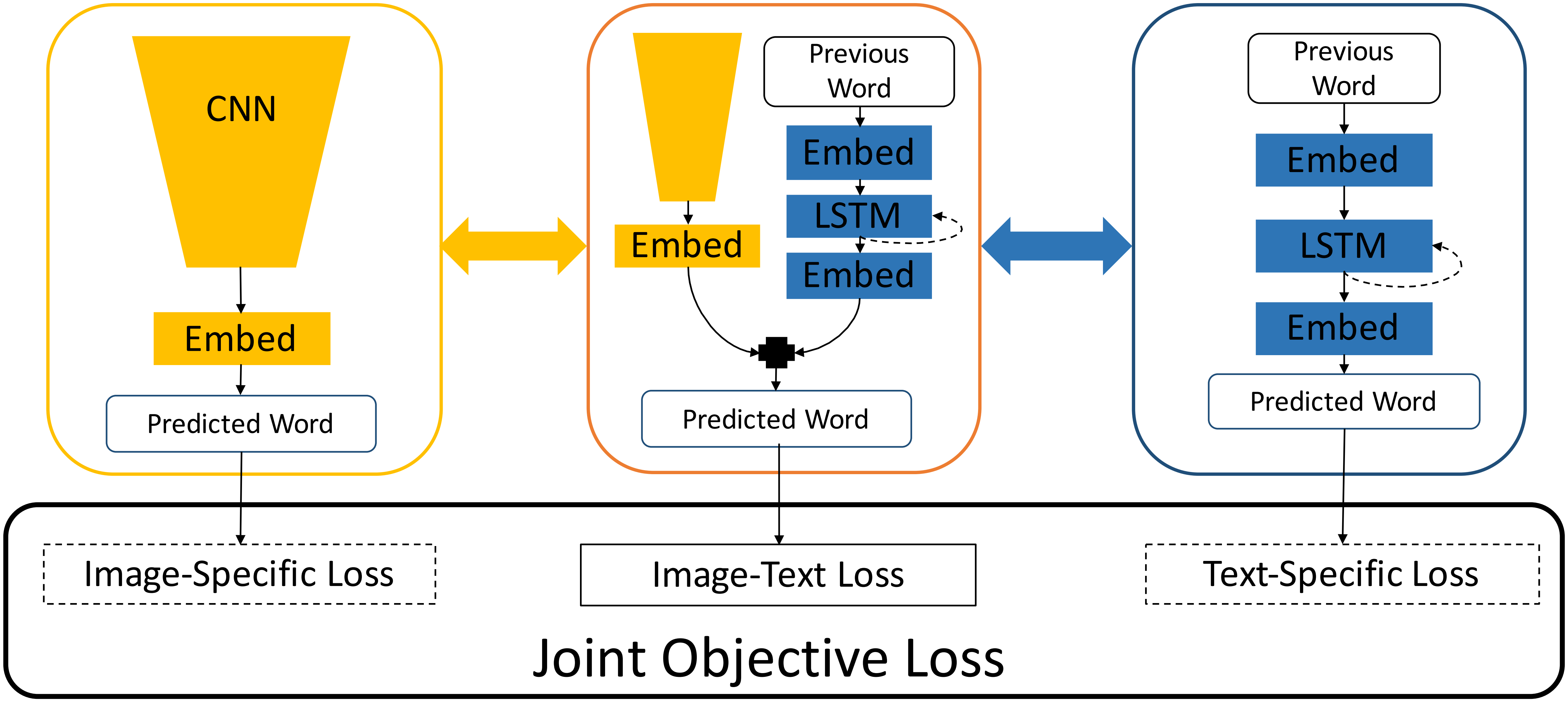
DCC Issue: Transfer Mechanism

A man is playing ***racket*** *on a* ***racket****.*

NOC Solution: Semantic Embedding



Training


CNN
Embed
Predicted Word
Previous Word
Embed
LSTM
Embed
Embed
Predicted Word
Previous Word
Embed
LSTM
Embed
Predicted Word
Image-Specific Loss
Image-Text Loss
Text-Specific Loss
Joint Objective Loss

F1 Scores for NOC and DCC

	Bottle	Bus	Couch	Microwave	Pizza	Racket	Suitcase	Zebra	Average
DCC	4.63	29.79	**45.87**	**28.09**	64.59	52.24	13.16	79.88	39.78
NOC	**19.02**	**69.34**	33.25	26.46	**69.16**	**62.45**	**34.65**	**89.78**	**50.51**

Ablation: Auxiliary Objective

Contributing Factor	Glove	LM Pretrain	Image Pretrain	Auxiliary Objective	Meteor	F1
Pretraining & Glove	**X**	**X**	**X**		19.80	25.38
Fix Image Model	**X**	**X**	Fixed		18.91	39.70
All	**X**	**X**	**X**	**X**	**20.69**	**50.51**

Ablation: Glove Embedding

Contributing Factor	Glove	LM Pretrain	Image Pretrain	Auxiliary Objective	Meteor	F1
Auxiliary Objective			**X**	**X**	15.78	14.41
Glove	**X**		**X**	**X**	19.69	47.02
All	**X**	**X**	**X**	**X**	**20.69**	**50.51**

Training with Outside Data

Image Data	Text Data	Meteor	F1
MSCOCO	MSCOCO	20.69	50.51
MSCOCO	WebCorpus	19.15	41.74
ImageNet	WebCorpus	17.55	36.50

Describing ImageNet

A **otter** is sitting on a rock in the sun.

A large **flounder** is resting on a rock.

A table with a plate of **sashimi** and vegetables.

A large **glacier** with a mountain in the background.

A man is standing on a beach holding a **snapper**.

A group of people standing around a large white **warship**.

Errors

A **chainsaw** is sitting on a **chainsaw** near a **chainsaw**.

A **volcano** view of a **volcano** in the sun.

Compositional Captioner

A ***anteater*** *is standing in the grass.*

Unpaired Image Data

Paired Image-Sentence Data

Unpaired Text Data

Existing Methods

LRCN: *A black* ***bear*** *is standing in the grass.*

CaptionBot: *A* ***bear*** *that is eating some grass.*

Our Team:

Lisa Anne Hendricks

Subhashini Venugopalan

Marcus Rohrbach

Raymond Mooney

Kate Saenko

Trevor Darrell