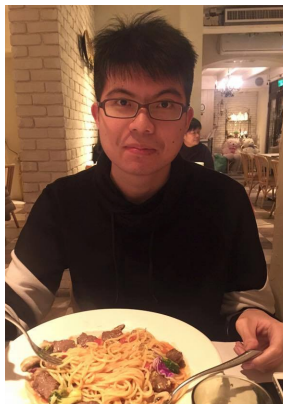
Show, Adapt and Tell:
Adversarial Training of Cross-domain Image Captioner



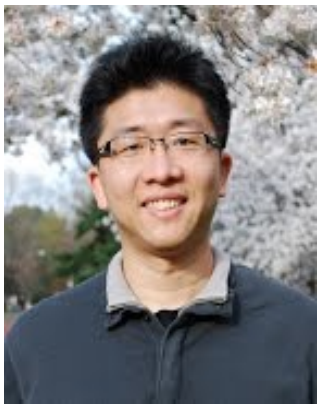
Tseng-Hung Chen
NTHU EE

Yuan-Hong Liao
NTHU EE

Ching-Yao Chuang
NTHU EE

Wan-Ting Hsu
NTHU EE

Jianlong Fu
MSRA



Min Sun
NTHU EE

Outline

- Motivation
- Introduction
- Cross-domain Image Captioning
- Critic-based Planning
- Experiment
- Conclusion

Motivation

paired data in domain1

A young girl riding a surfboard in the ocean.

A hummingbird close to a flower trying to eat.

A dog laying on the grass with a frisbee.

Cross-domain Image Captioner

paired data in domain2

This bird has wings that are brown and has red eyes.

A small bird with orange flank and a long thin black bill.

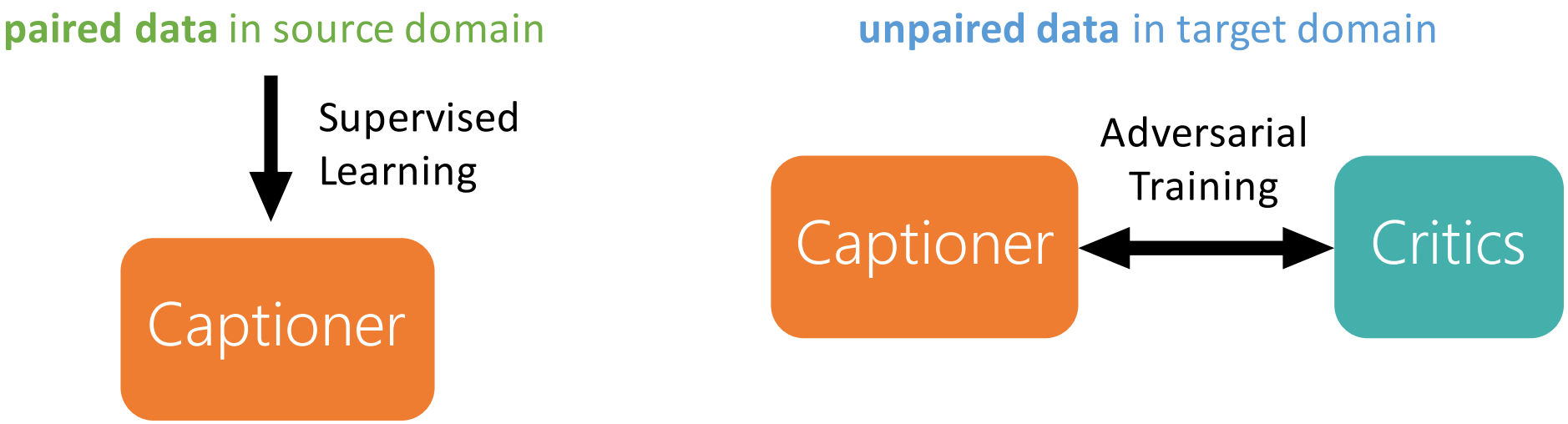
high cost!

Outline

- Motivation
- Introduction
- Cross-domain Image Captioning
- Critic-based Planning
- Experiment
- Conclusion

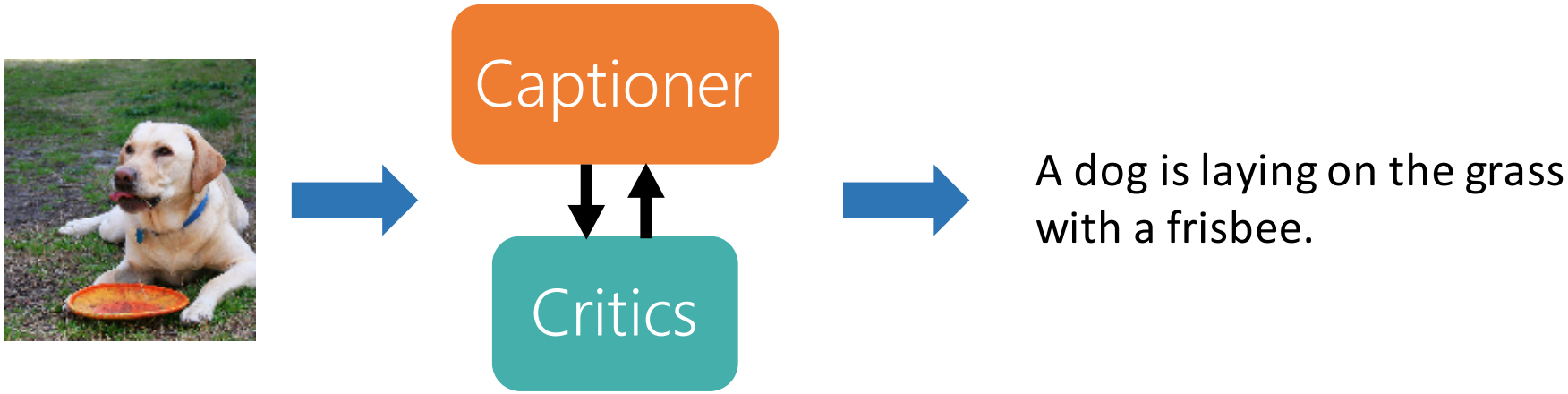
Introduction

- We propose a novel adversarial training procedure for cross-domain captioner. It utilizes critics to capture the distribution of image and sentence in the target domain.



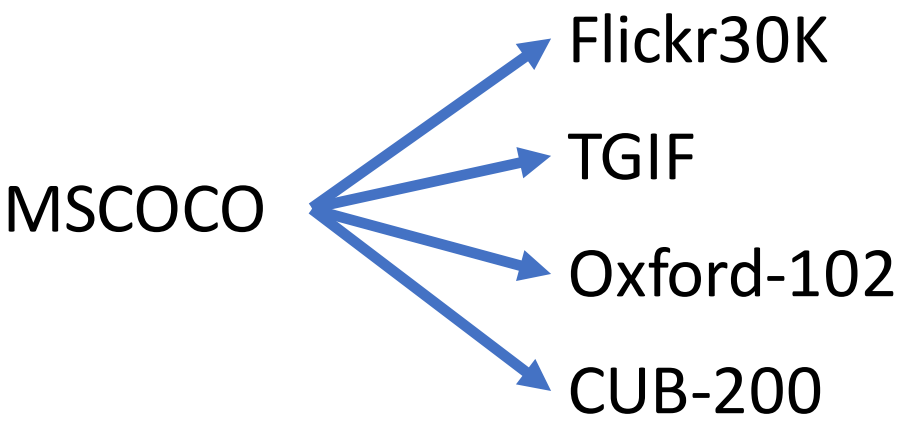
Introduction

- We propose to utilize the knowledge of critics during inference to further improve the performance.



Introduction

- Our method achieves significant improvement on four publicly available datasets compared to a captioner trained only on the source domain.



Outline

- Motivation
- Introduction
- Cross-domain Image Captioner
- Critic-based Planning
- Experiment
- Conclusion

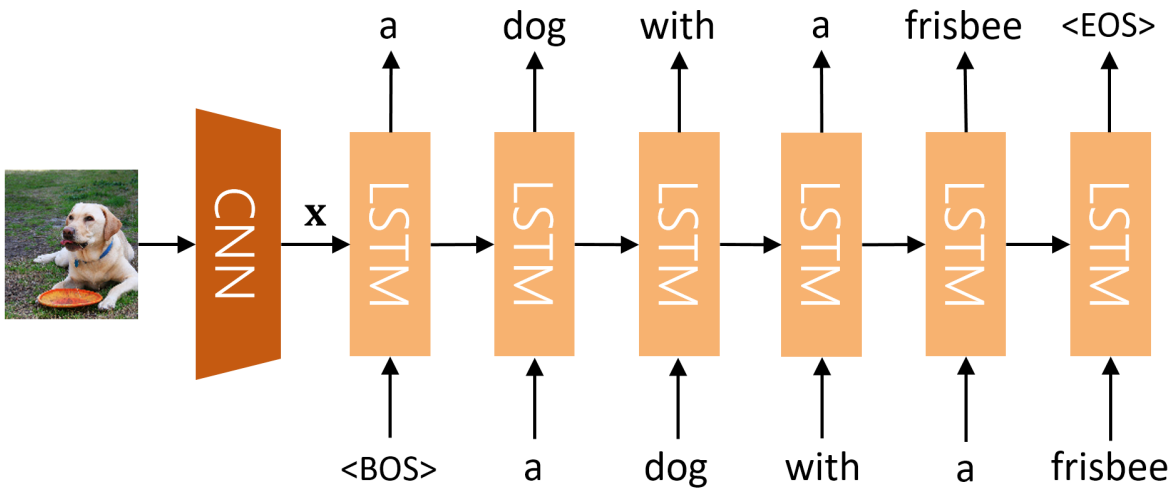
Approach

1. pre-train captioner in source domain

source paired data

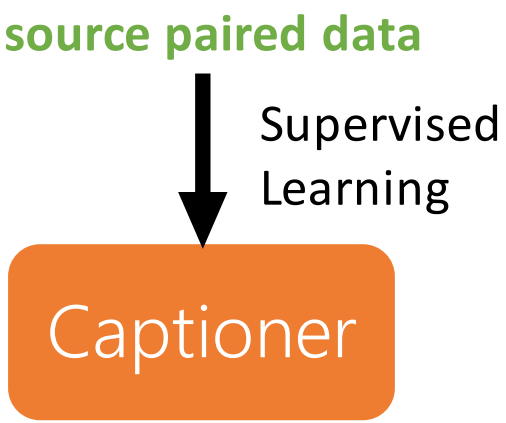
Supervised Learning

Captioner

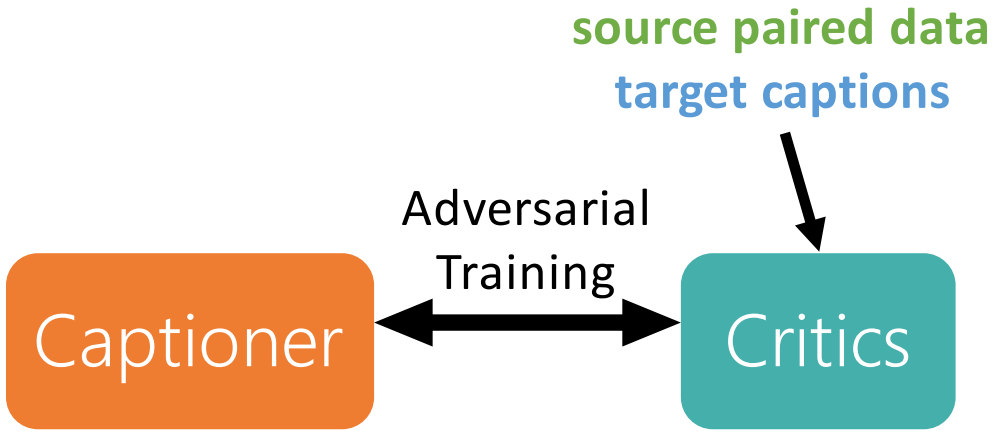


Approach

1. pre-train captioner in source domain



2. adversarial training (critic guide captioner) in target domain



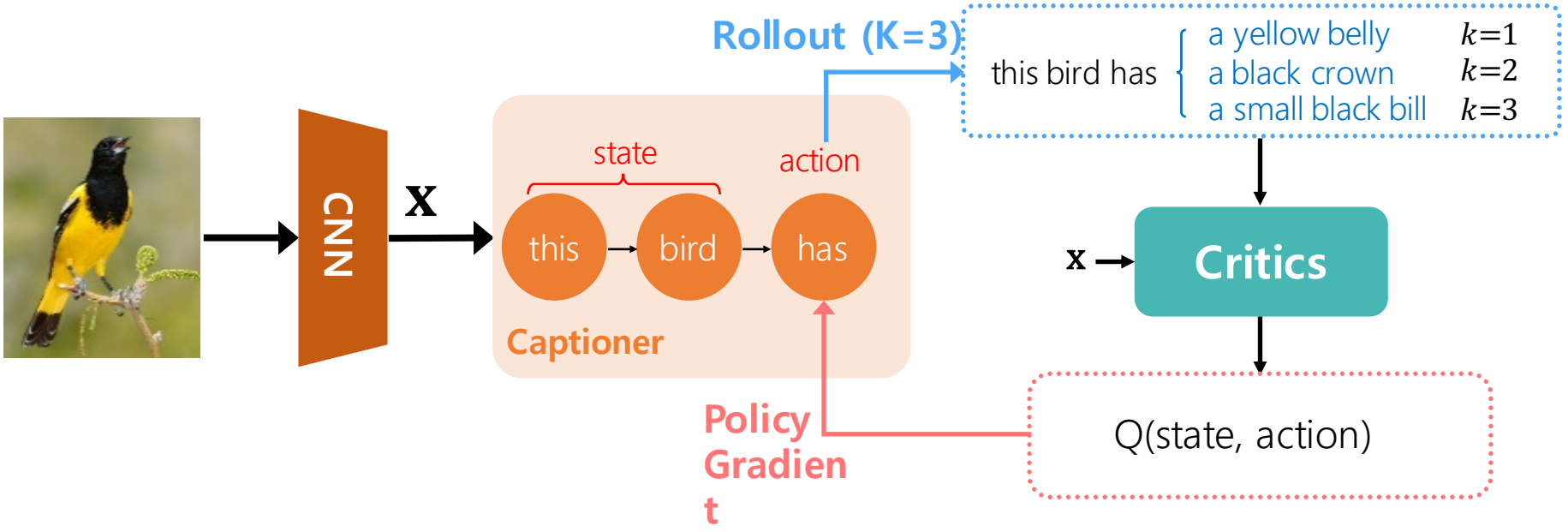
Adversarial Training

We propose a novel adversarial training procedure to leverage unpaired images and sentences.

Captioner is treated as an agent taking sequential actions and **Critic** evaluates agent's behavior.

During training, **Critic** and **Captioner** act as adversaries – **Captioner** aims to generate indistinguishable captions, whereas **Critic** aims at distinguishing them.

Captioner as an Agent



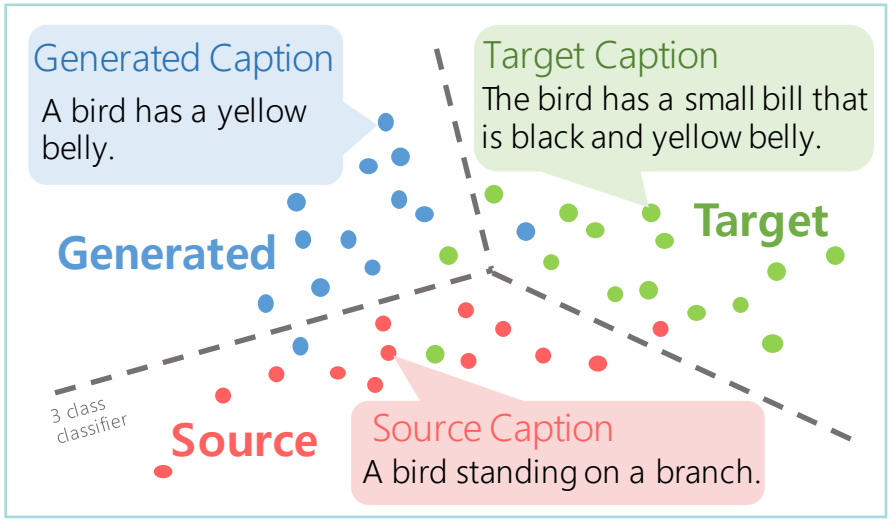
Critics

For cross-domain image captioning, a good caption needs to satisfy two criteria:

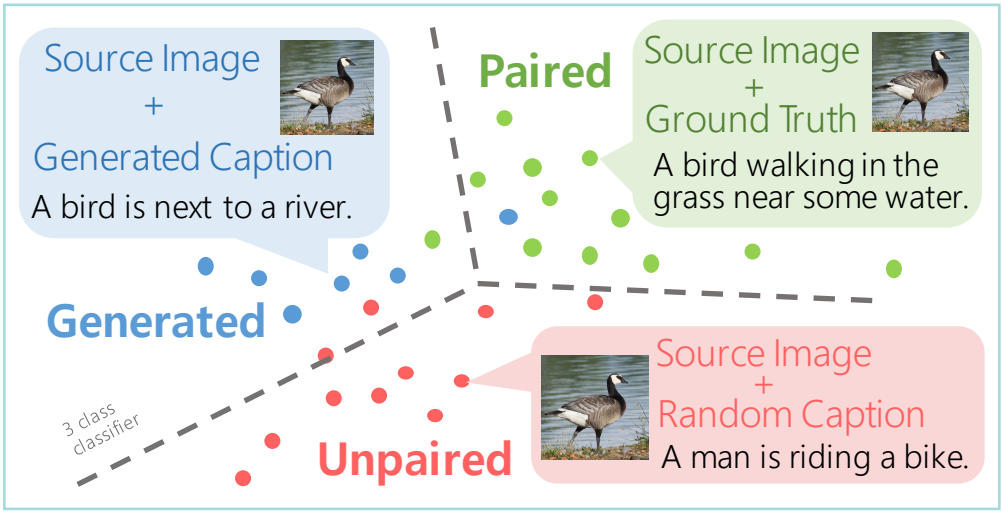
(1) the generated sentence resembles the sentence drawn from the target domain.

(2) the generated sentence is relevant to the input image.

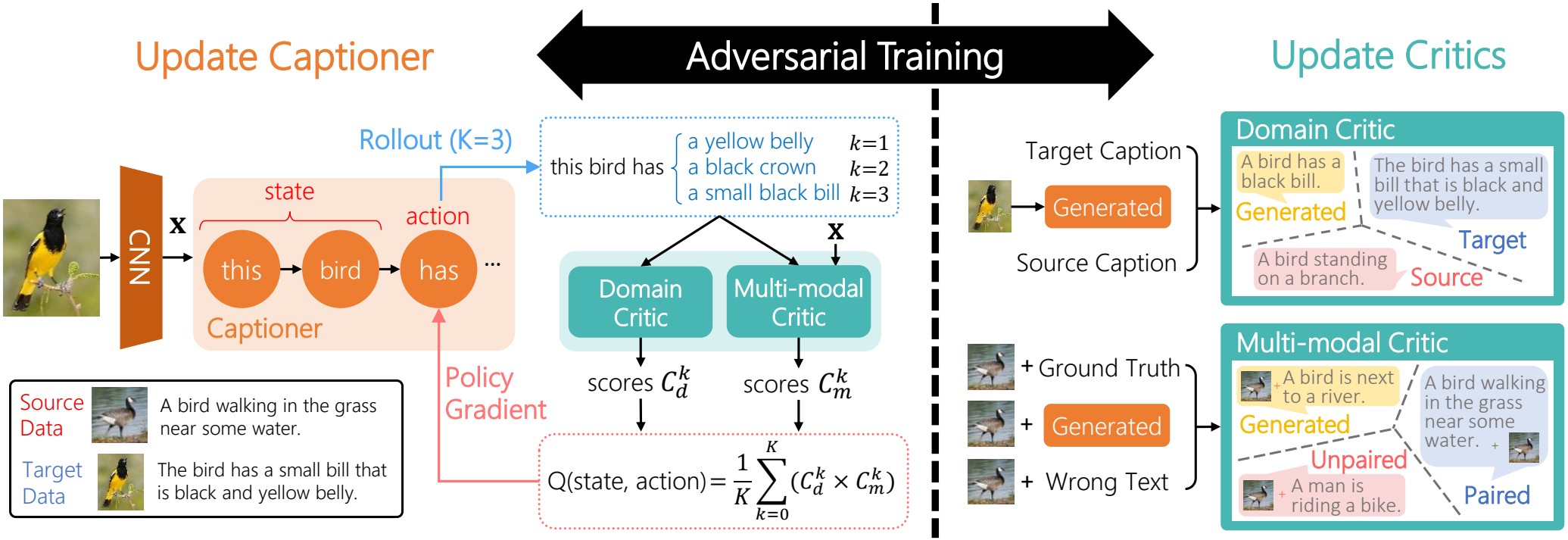
Domain Critic



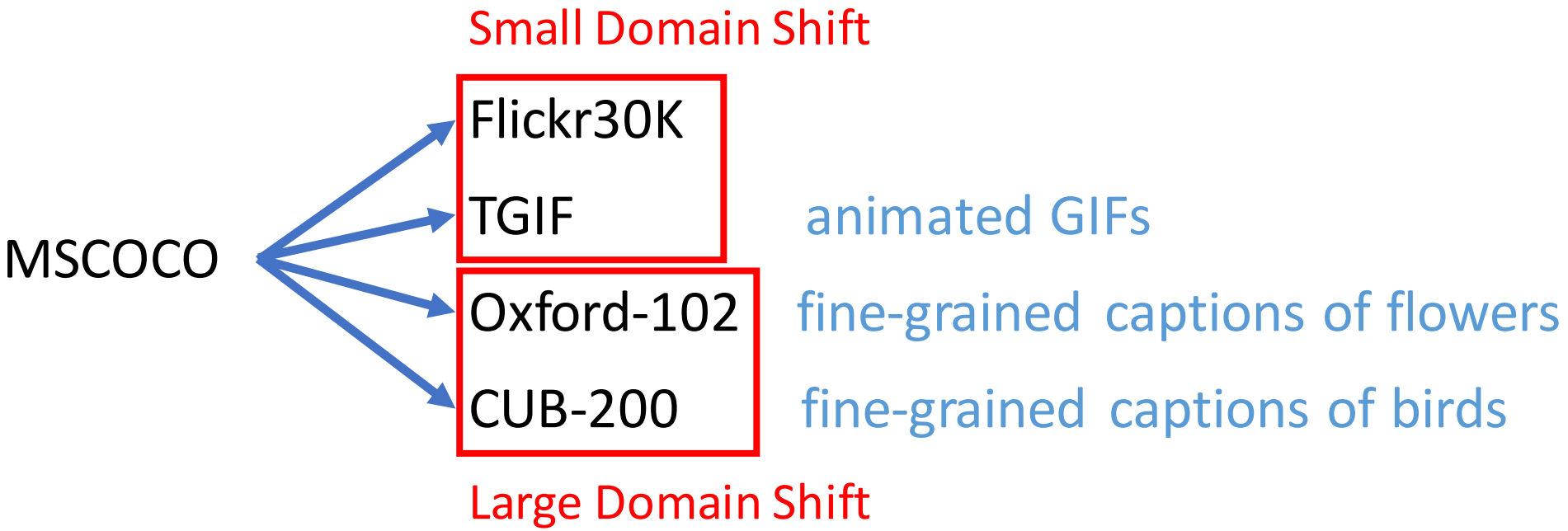
Multi-modal Critic



Adversarial Training



MSCOCO Adapt to Four Datasets



Domain Shift across Datasets

Word-level Distribution

(a) MSCOCO

Ex: A man in black shirt is next to the river.

(b) Oxford-102

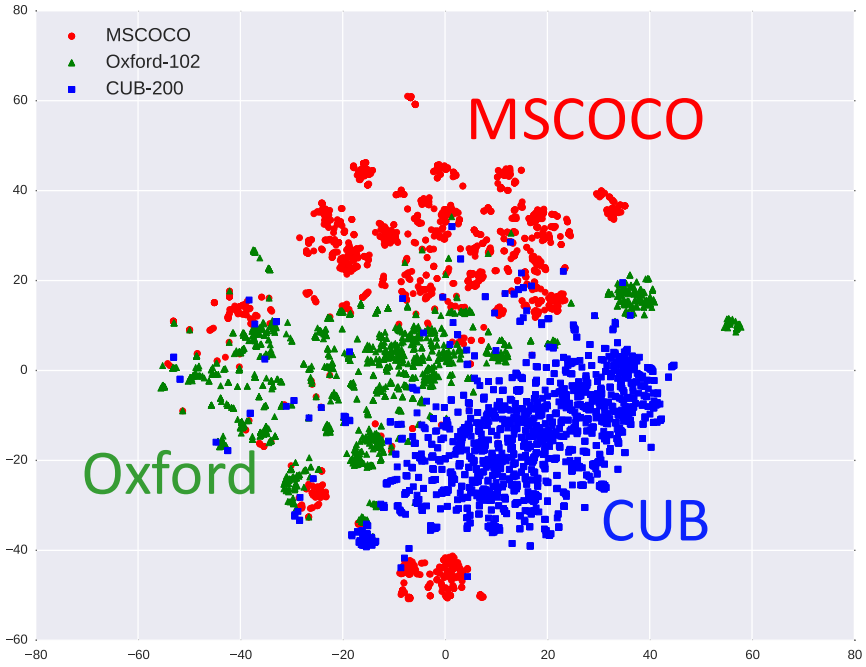
Ex: This flower has yellow petals and red center.

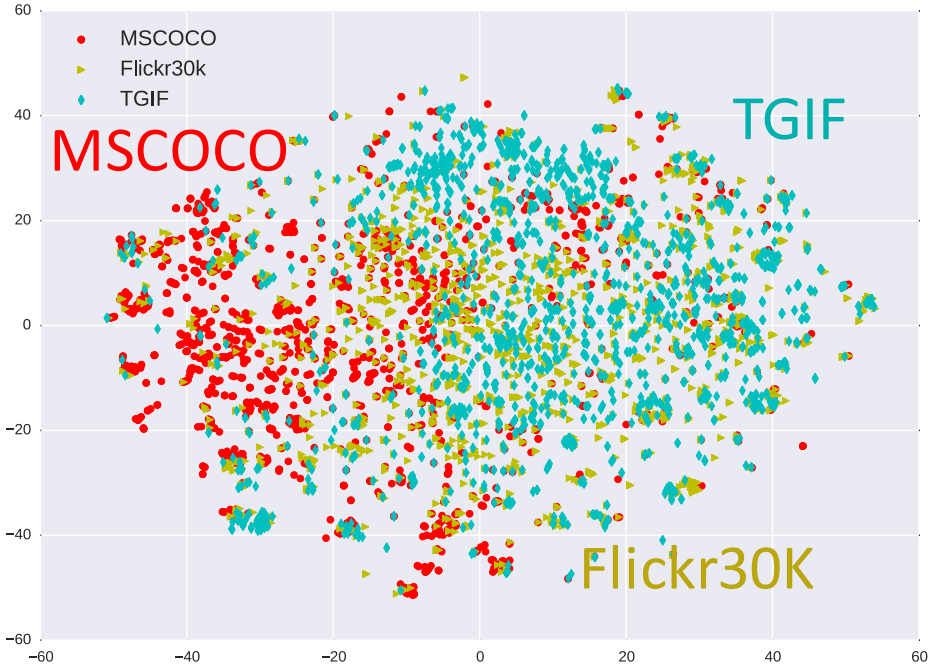
(c) CUB-200

Ex: A small bird with orange flank and a long thin black bill.

Domain Shift across Datasets

Sentence-level Embedding





Results on Four Datasets

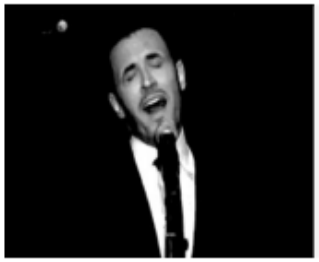
Method	Target (test)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	ROUGE	CIDEr	SPICE
Source Pre-trained	CUB-200	50.8	28.3	13.9	6.1	12.9	33	3	4.6
DCC	CUB-200	68.6	47.3	31.4	21.4	23.8	46.4	11.9	11.1
Ours	CUB-200	**91.4**	**73.1**	**51.9**	**32.8**	**27.6**	**58.6**	**24.8**	**13.2**
Fine-tuning	CUB-200	91.3	80.2	69.2	59	36.1	69.7	61.1	17.9
Source Pre-trained	Oxford-102	48.3	21.6	6.2	1.3	10.5	25.8	3.1	4.4
DCC	Oxford-102	51	33.8	24.1	16.7	21.5	38.3	6	9.8
Ours	Oxford-102	**85.6**	**76.9**	**67.4**	**60.5**	**36.4**	**72.1**	**29.3**	**17.9**
Fine-tuning	Oxford-102	87.5	80.1	72.8	66.3	40	75.6	36.3	18.5
Source Pre-trained	TGIF	41.6	23.3	12.6	7	12.7	32.7	14.7	8.5
DCC	TGIF	34.6	17.5	9.3	4.1	11.8	29.5	7.1	7.3
Ours	TGIF	**47.5**	**29.2**	**17.9**	**10.3**	**14.5**	**37**	**22.2**	**10.6**
Fine-tuning	TGIF	51.1	32.2	20.2	11.8	16.2	39.2	29.8	12.1
Source Pre-trained	Flickr30k	57.3	36.2	21.9	13.3	15.1	38.8	25.3	8.6
DCC	Flickr30k	54.3	34.6	21.8	13.8	16.1	38.8	27.7	9.7
Ours	Flickr30k	**62.1**	**41.7**	**27.6**	**17.9**	**16.7**	**42.1**	**32.6**	**9.9**
Fine-tuning	Flickr30k	59.8	41	27.5	18.3	18	42.9	35.9	11.5

Results on Small Domain Shift

MSCOCO ➔ TGIF

Before: A cat is standing in a room with a cat.

After: A cat is playing with a toy in a room.



Before: A man in a black shirt and a tie.

After: A man in a suit is singing into a microphone.

MSCOCO ➔ Flickr30K

Before: A young baseball player is a ball in the field.

After: A young baseball player is sliding into a base.

Before: A boy in a field playing with a frisbee.

After: A young boy playing with a soccer ball in a field.

Results on Large Domain Shift

MSCOCO ➔ CUB-200

Before: A bird is standing on a table with flowers.

After: A small bird with a white belly and a black head.

Before: A red bird sitting on a tree branch.

After: This is a red bird with a black wing and a small beak.

MSCOCO ➔ Oxford-102

Before: A white flower in a vase on a table.

After: This flower has petals that are pink and has a yellow center.

Before: A yellow flower is in a clear vase.

After: This flower has petals that are yellow and has red lines.

Results on Critic-based Planning

Method	Bleu-4	Meteor	ROUGE	CIDEr-D
MSCOCO → CUB-200				
Greedy Search	32.8	27.6	58.6	24.8
Beam Search	33.1	27.5	58.3	26.2
Planning	**35.2**	27.4	58.5	**29.3**
MSCOCO → Oxford-102				
Greedy Search	60.5	36.4	72.1	**29.3**
Beam Search	60.3	36.3	72	28.3
Planning	**62.4**	**36.6**	**72.6**	24.9
MSCOCO → TGIF				
Greedy Search	10.3	14.5	37	22.2
Beam Search	10.5	14.2	36.7	22.6
Planning	10.3	14.4	37	21.9
MSCOCO → Flickr30k				
Greedy Search	17.5	16.4	41.9	32.2
Beam Search	18.2	16.4	42.1	33.3
Planning	17.3	16.5	41.7	32.3

Greedy Search:
A small bird with a long beak and a long beak.

Critic-based planning:
A black and white bird with a long beak.