Adversarial Representation Learning for Text-to-Image Matching

Nikolaos Sarafianos, Xiang Xu, Ioannis A. Kakadiaris
{nsarafianos, xxu21, ioannisk}@uh.edu

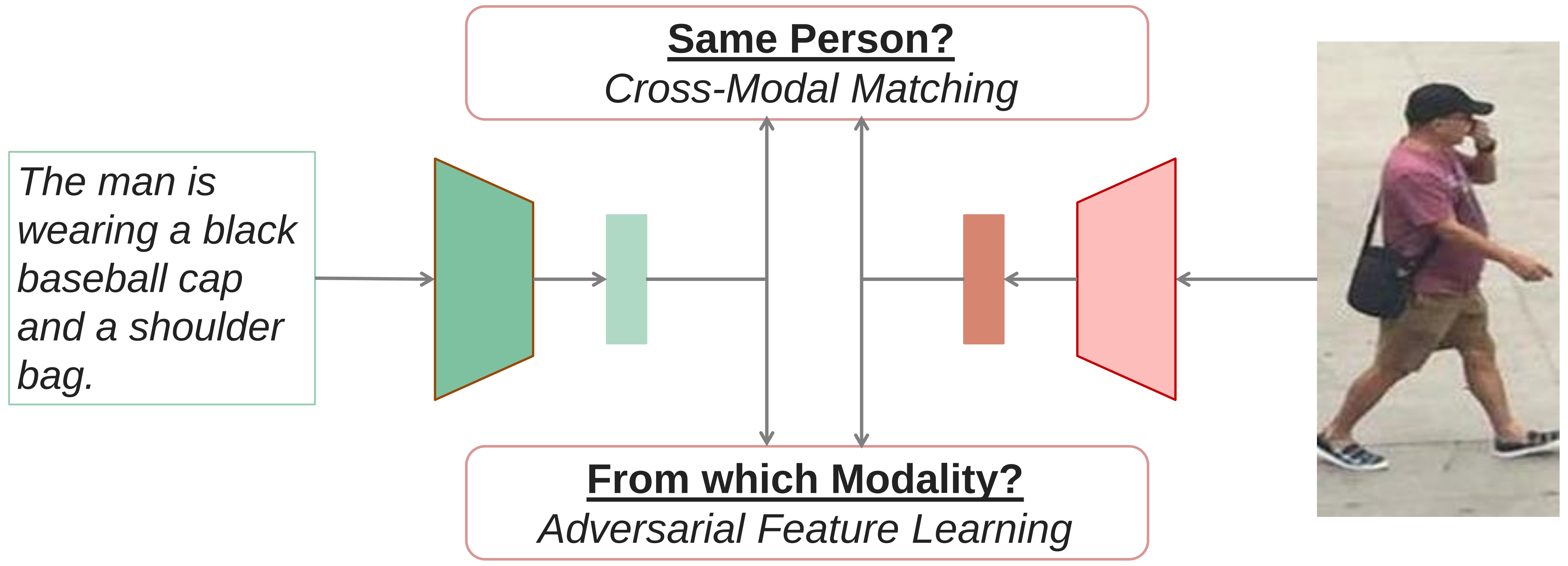
ICCV 2019
Seoul, Korea

Motivation

Problem Statement: Given a textual description retrieve the most relevant images

Objectives:

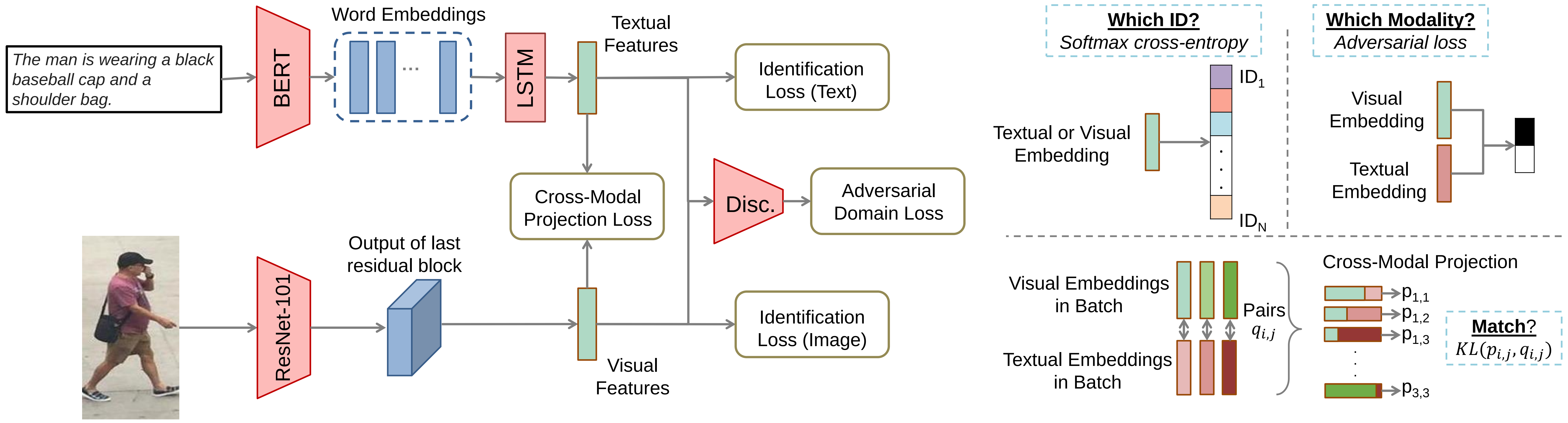
- Match the distributions of the features that belong to the same identity
- Learn modality invariant representations


Same Person?
Cross-Modal Matching

The man is wearing a black baseball cap and a shoulder bag.

From which Modality?
Adversarial Feature Learning

Method



Contributions:

1. Introduce an adversarial representation learning framework that brings the features from both modalities "close-to-each-other"
2. Demonstrate that BERT can result in more discriminative word representations suitable for cross-modal matching

Quantitative Results

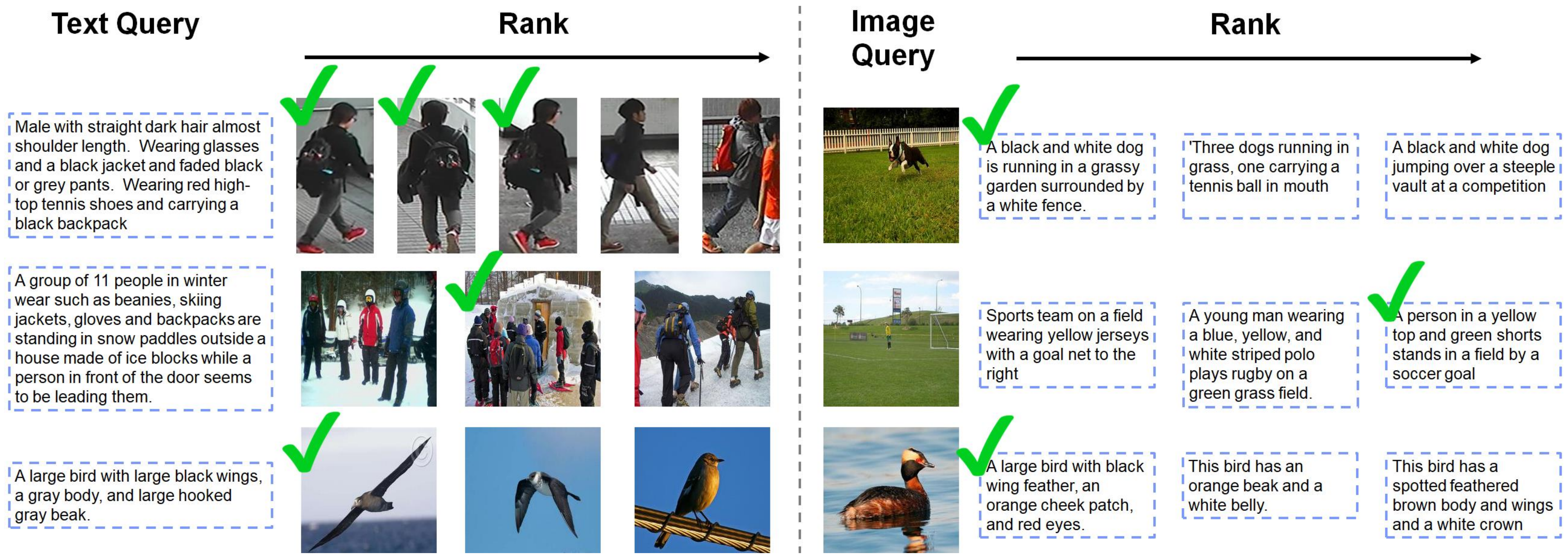
Cross-Modal Retrieval results on the Flickr-30K dataset

Method	Image-to-Text		Text-to-Image	
	Rank-1	Rank-10	Rank-1	Rank-10
DAN	55.0	89.0	39.4	79.1
NAR	55.1	89.6	39.4	79.9
VSE++	52.9	87.2	39.6	79.5
SCO	55.5	89.3	41.1	80.1
GXN	**56.8**	**89.6**	41.5	80.1
TIMAM	53.1	87.6	**42.6**	**81.9**

Text-to-Image Retrieval Ablation Study on the CUHK-PEDES dataset

L_I	L_M	BERT	Adv. Learning	Rank-1
✓				40.1
	✓			44.9
✓	✓			49.8
✓	✓		✓	51.3
✓	✓	✓		52.9
✓	✓	✓	✓	54.5

Qualitative Results



Male with straight dark hair almost shoulder length. Wearing glasses and a black jacket and faded black or grey pants. Wearing red high-top tennis shoes and carrying a black backpack

A group of 11 people in winter wear such as beanies, skiing jackets, gloves and backpacks are standing in snow paddles outside a house made of ice blocks while a person in front of the door seems to be leading them.

A large bird with large black wings, a gray body, and large hooked gray beak.

A black and white dog is running in a grassy garden surrounded by a white fence.

'Three dogs running in grass, one carrying a tennis ball in mouth

A black and white dog jumping over a steeple vault at a competition

Sports team on a field wearing yellow jerseys with a goal net to the right

A young man wearing a blue, yellow, and white striped polo plays rugby on a green grass field.

A person in a yellow top and green shorts stands in a field by a soccer goal

A large bird with black wing feather, an orange cheek patch, and red eyes.

This bird has an orange beak and a white belly.

This bird has a spotted feathered brown body and wings and a white crown

Key Takeaways

- **Adversarial learning is well-suited for cross-modal matching:** Observed 2% to 5% improvements in terms of rank-1 accuracy over the previous best-performing techniques
- **Pre-trained language models can successfully be applied to cross-modal matching**: Observed 3% to 5% improvements when features are learned in this manner

UNIVERSITY of HOUSTON | CBL

Changing the way ~~people~~ computers look at ~~computers~~ people

Acknowledgements: This work has been funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.