Millennium Institute
Foundational
Research on Data

impresee

Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding

Jesus Perez-Martin, Benjamin Bustos, Jorge Pérez, Juan M. Barrios

TRECVID 2020 Workshop
2020-12-08

Problem: Video Captioning

Possible captions:

1. the men are fighting using martial arts.
2. men are doing martial arts.
3. the men are doing martial arts together in the field.

Possible captions:

1. a woman is applying makeup on her face
2. a young lady is doing makeup on her face
3. a girl applying blush on her face

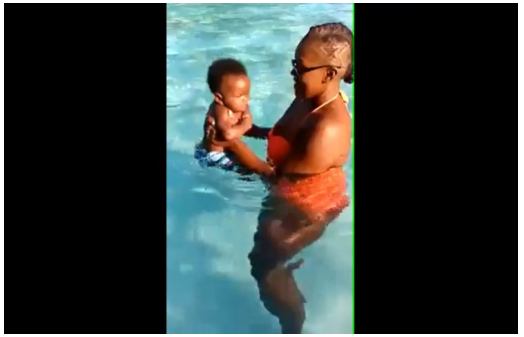
Syntactic Patterns in Video Descriptions

object1 **object2** **action** **object3**

275. a woman and a child swim in a pool...
560. … a man and a dog ... sitting at the top of the tree...
690. a man and a pig are walking along a sidewalk at daytime.
1100. a woman and a man driving in a car
1354. a man and a woman ... hold a microphone
1677. a man and a woman sitting in a radio studio are shown.
1883. a man and a young girl riding a marry-go-round
2041. A Spanish-speaking man and a Spanish-speaking woman argue a TV split screen...
2215. A big dog and a small dog sharing a bone.
2632. A young man and a woman brush their teeth in a bathroom.
2769. A young man and a woman are kissing, in a room.
2785. A woman and a boy are watching a video...

Syntactic Patterns in Video Descriptions

object1 **object2** **action** **object3**



275. a woman and a child swim in a pool...
560. … a man and a dog ... sitting at the top of the tree...
690. a man and a pig are walking along a sidewalk at daytime.
1100. a woman and a man driving in a car
1354. a man and a woman ... hold a microphone
1677. a man and a woman sitting in a radio studio are shown.
1883. a man and a young girl riding a marry-go-round
2041. A Spanish-speaking man and a Spanish-speaking woman argue a TV split screen...
2215. A big dog and a small dog sharing a bone.
2632. A young man and a woman brush their teeth in a bathroom.
2769. A young man and a woman are kissing, in a room.
2785. A woman and a boy are watching a video...

Syntactic Patterns in Video Descriptions

object1 **object2** **action** **object3**

275. a woman and a child swim in a pool...
560. … a man and a dog ... sitting at the top of the tree...
690. a man and a pig are walking along a sidewalk at daytime.
1100. a woman and a man driving in a car
1354. a man and a woman ... hold a microphone
1677. a man and a woman sitting in a radio studio are shown.
1883. a man and a young girl riding a marry-go-round
2041. A Spanish-speaking man and a Spanish-speaking woman argue a TV split screen...
2215. A big dog and a small dog sharing a bone.
2632. A young man and a woman brush their teeth in a bathroom.
2769. A young man and a woman are kissing, in a room.
2785. A woman and a boy are watching a video...

Syntactic Patterns in Video Descriptions

object1 **object2** **action** **object3**

275. a woman and a child swim in a pool...
560. … a man and a dog ... sitting at the top of the tree...
690. a man and a pig are walking along a sidewalk at daytime.
1100. a woman and a man driving in a car
1354. a man and a woman ... hold a microphone
1677. a man and a woman sitting in a radio studio are shown.
1883. a man and a young girl riding a marry-go-round
2041. A Spanish-speaking man and a Spanish-speaking woman argue a TV split screen...
2215. A big dog and a small dog sharing a bone.
2632. A young man and a woman brush their teeth in a bathroom.
2769. A young man and a woman are kissing, in a room.
2785. A woman and a boy are watching a video...

Syntactic Patterns in Video Descriptions

object1 **object2** **action** **object3**

275. a woman and a child swim in a pool...
560. … a man and a dog ... sitting at the top of the tree...
690. a man and a pig are walking along a sidewalk at daytime.
1100. a woman and a man driving in a car
1354. a man and a woman ... hold a microphone
1677. a man and a woman sitting in a radio studio are shown.
1883. a man and a young girl riding a marry-go-round
2041. A Spanish-speaking man and a Spanish-speaking woman argue a TV split screen...
2215. A big dog and a small dog sharing a bone.
2632. A young man and a woman brush their teeth in a bathroom.
2769. A young man and a woman are kissing, in a room.
2785. A woman and a boy are watching a video...

Video Captioning with Visual-Syntactic Embedding

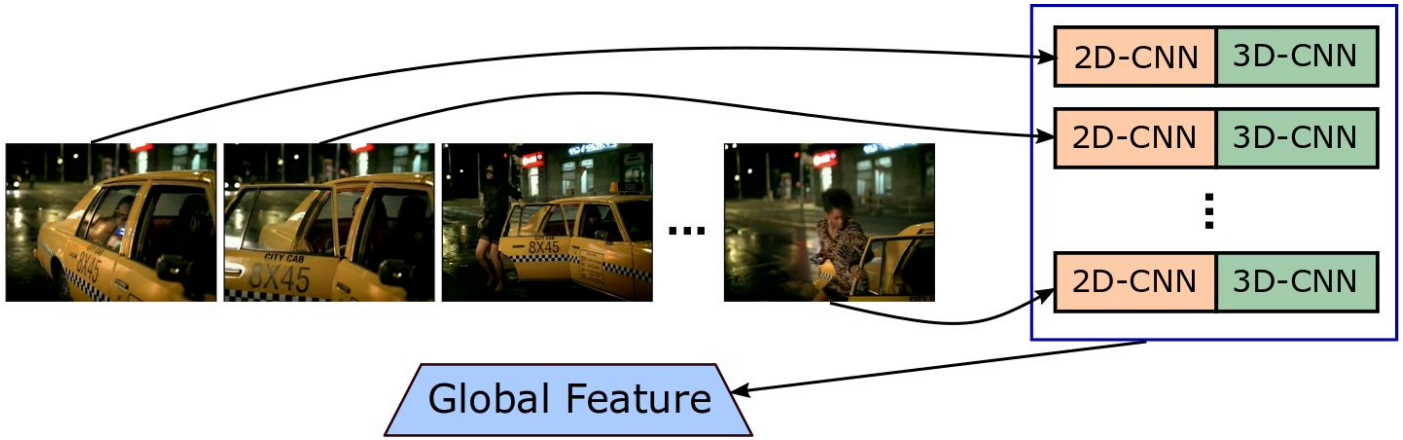
1. Cues about the syntactic structure of the video's descriptions can be directly extracted from a video

2. Existing models often produce syntactically incorrect sentences which harms their performance on standard datasets

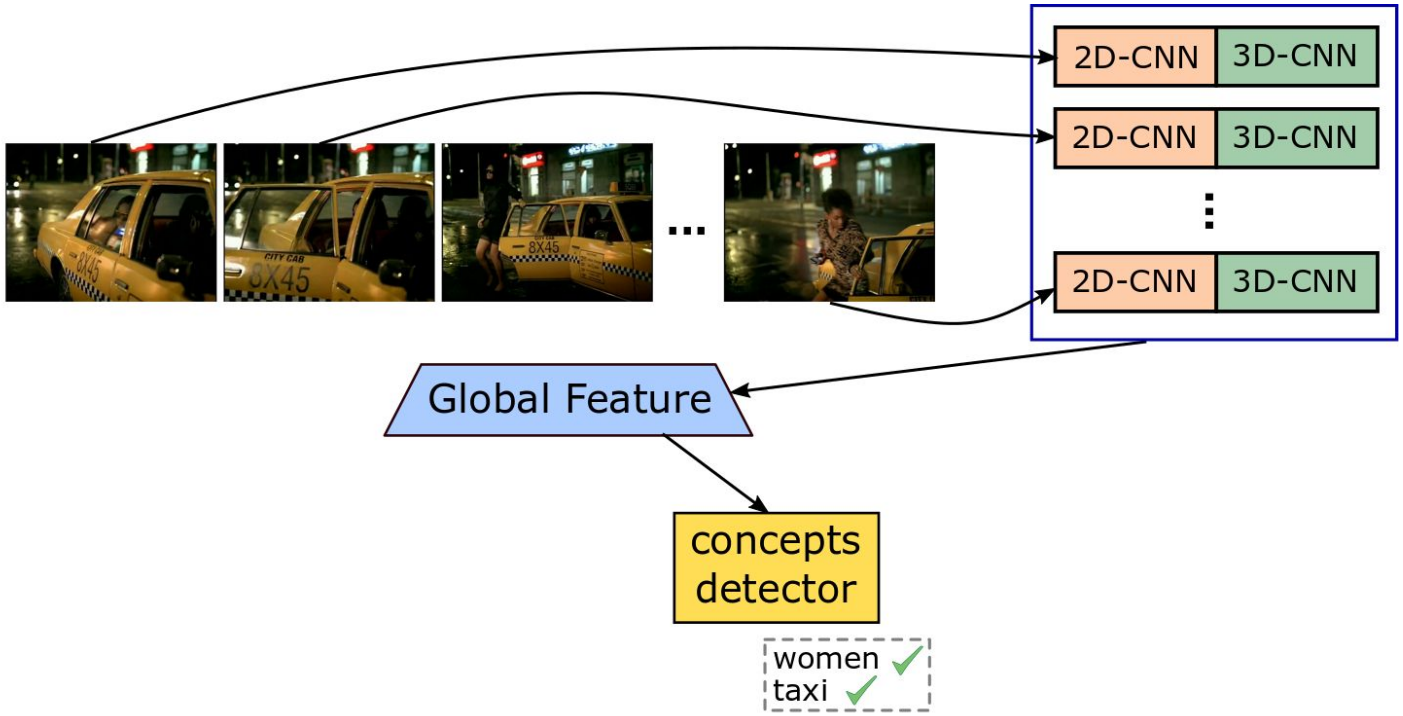
Model Overview

Model Overview

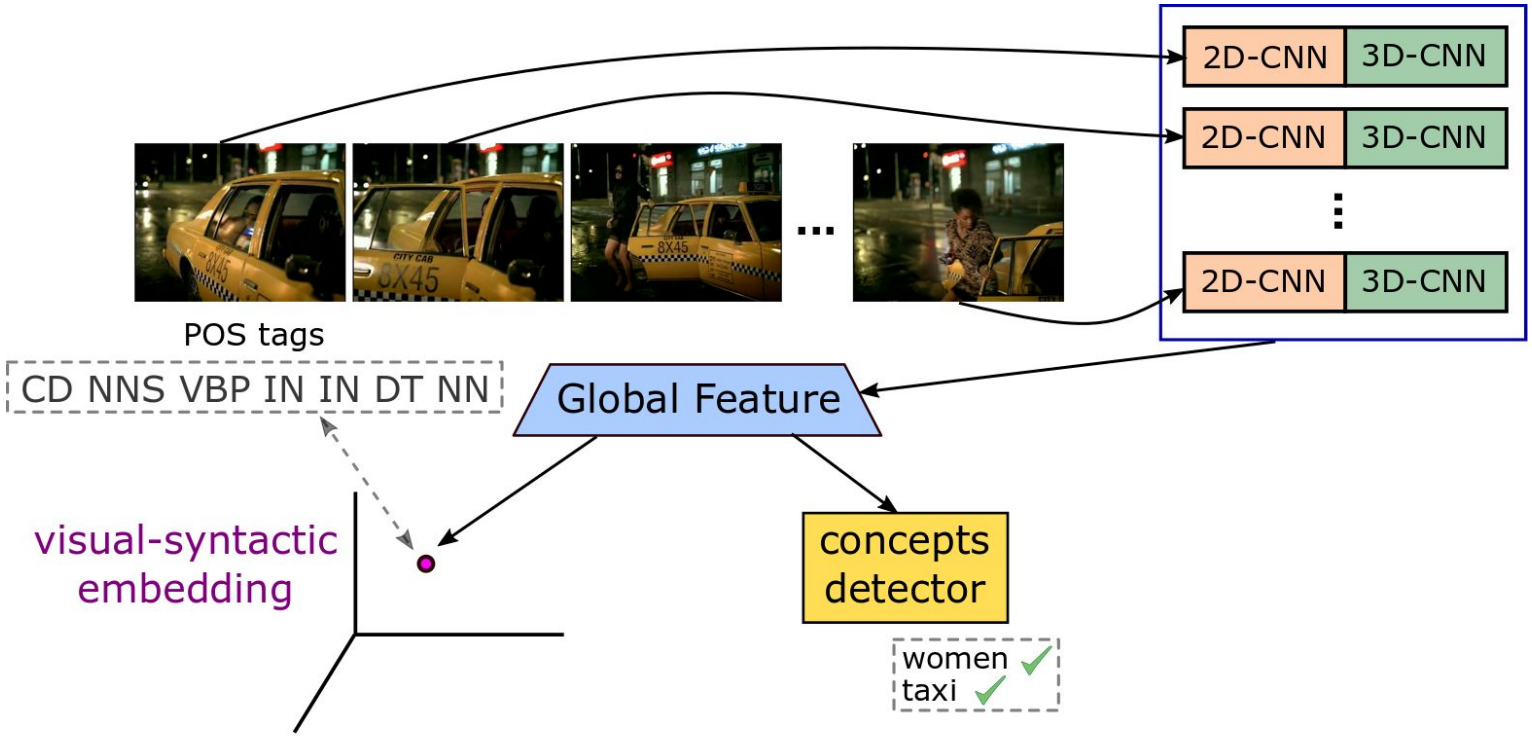
Model Overview



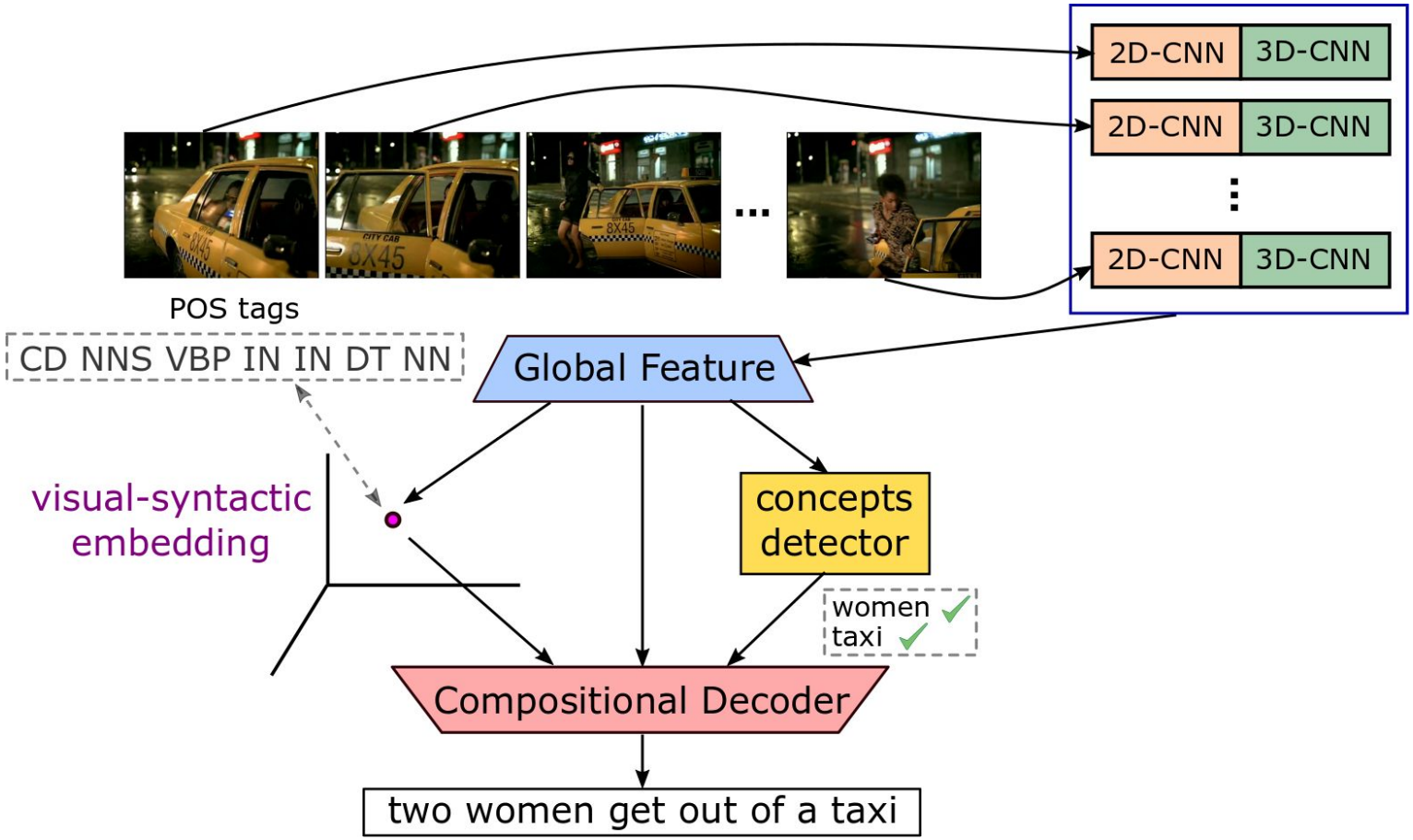
Model Overview



Model Overview



Model Overview



Visual-Syntactic Embedding

the men are fighting using martial arts

positive example

negative example

a woman is applying makeup on her face

Visual-Syntactic Embedding

the men are fighting using martial arts

positive example ↓ POS tagging

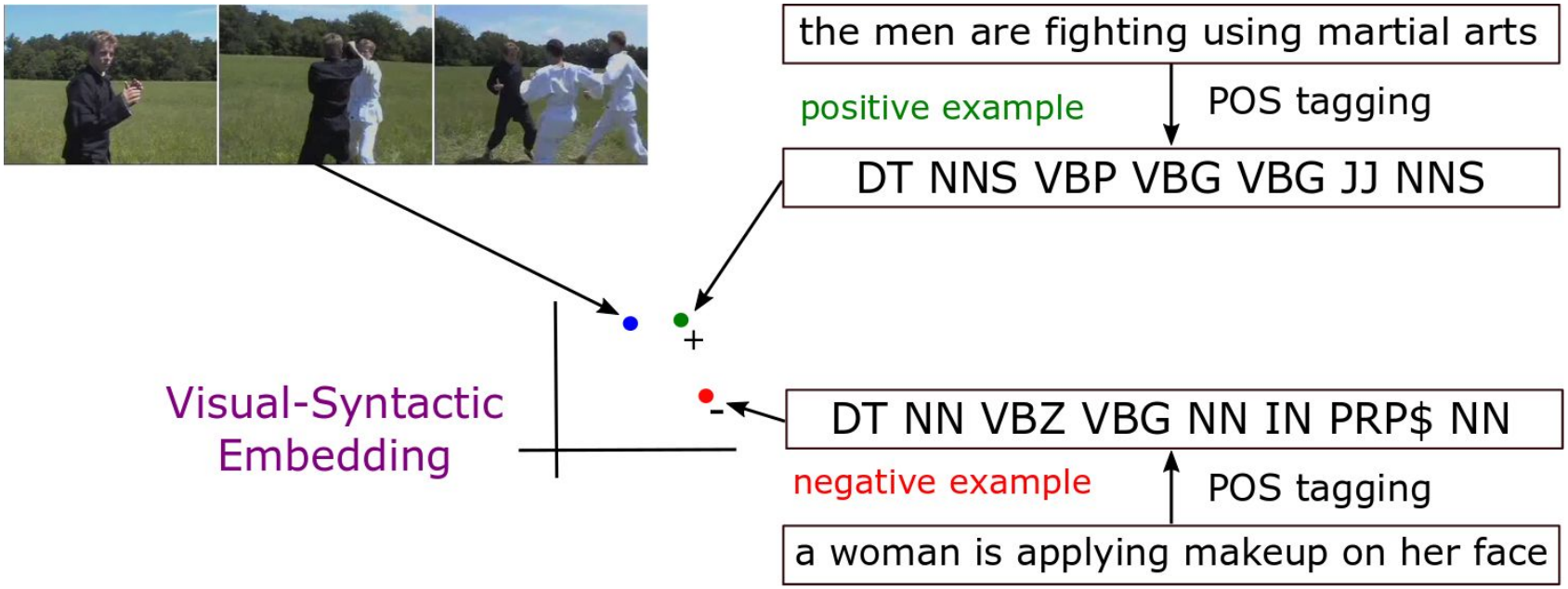
DT NNS VBP VBG VBG JJ NNS

DT NN VBZ VBG NN IN PRP$ NN

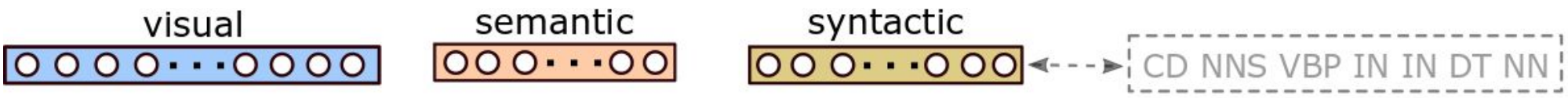
negative example ↑ POS tagging

a woman is applying makeup on her face

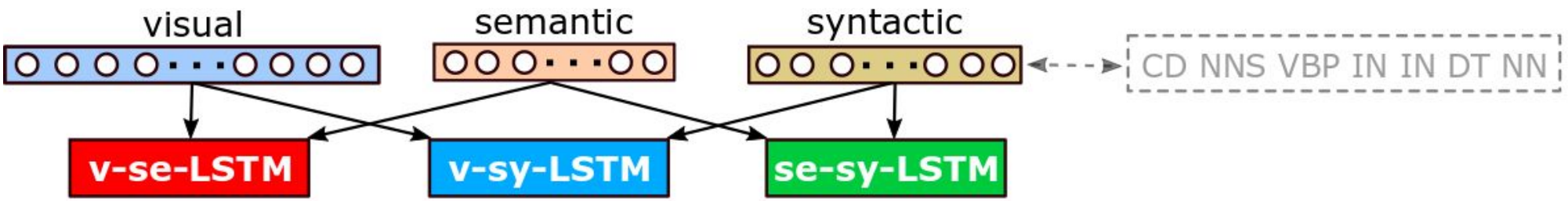
Visual-Syntactic Embedding



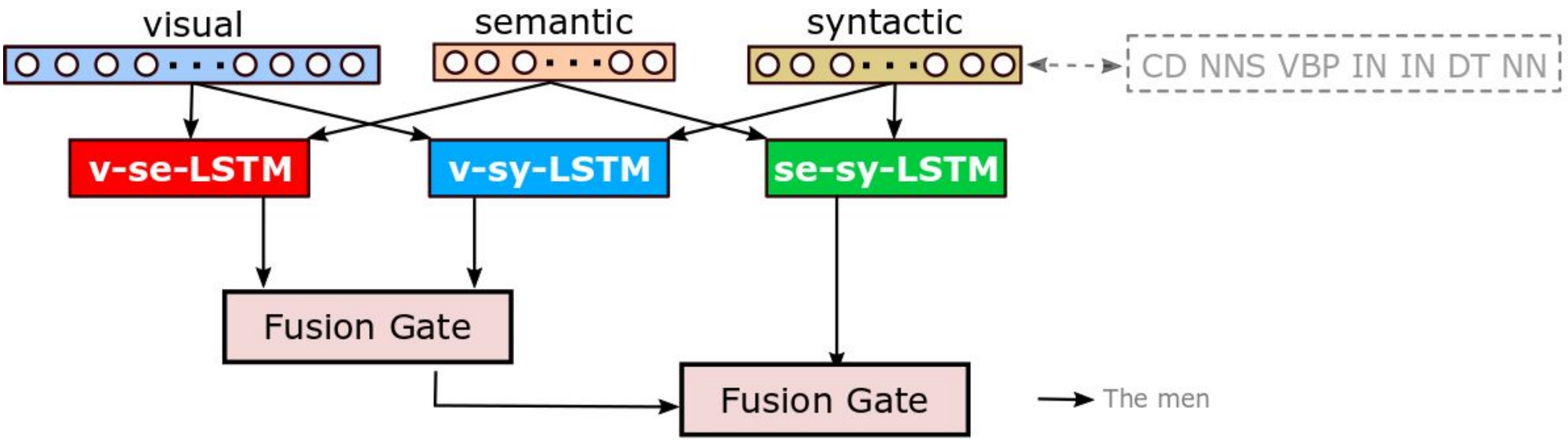
Compositional Decoder



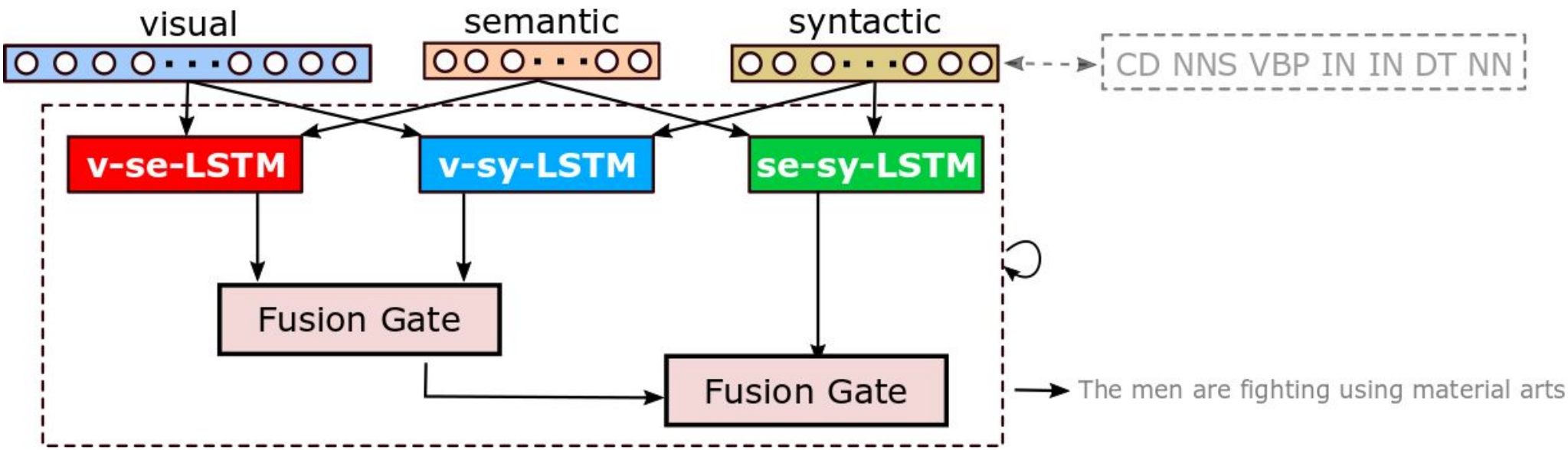
Compositional Decoder



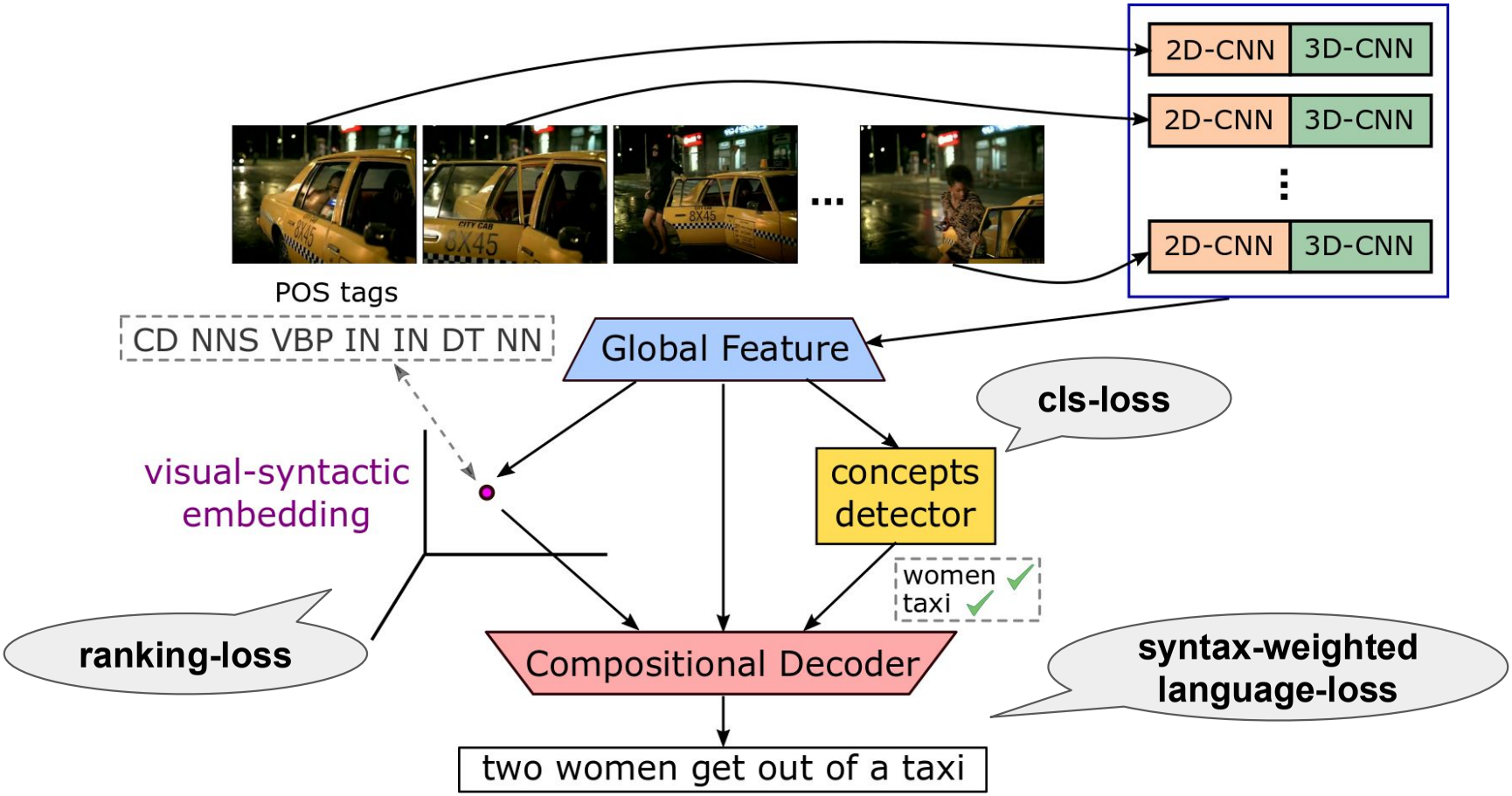
Compositional Decoder



Compositional Decoder



Training Process



Syntax-Weighted Loss

Given a video x, the ground-truth caption $y = (y_1, y_2, \ldots, y_L)$ of x, and the POS tagging t of the generated description, we define the weight

$$w = \max\left\{1, L^{\beta} - \left(\mathrm{dist}\left(\phi(\rho(x)), \omega(t)\right) + 1\right)^{\gamma}\right\},$$

and we minimize

$$\mathcal{L}_{\theta} = -\frac{1}{w}\sum_{i=1}^{L} \log p_{\theta}(y_i | y_{z<i})$$

Experiments - Datasets and Setup

MSVD

1,970 videos
- 1,200 train
- 100 validation
- 670 test

~ 40 captions per video

~ 6K words vocabulary

MSR-VTT

10,000 videos
- 6,512 train
- 498 validation
- 2,990 test

~ 20 captions per video

~ 14K words vocabulary

TRECVID 2020

9,185 videos
- 7,485 train
- 1,700 test

2 ~ 5 captions per video

~ 11K words vocabulary

Experiments - Datasets and Setup

MSVD

1,970 videos
- 1,200 train
- 100 validation
- 670 test

~ 40 captions per video

~ 6K words vocabulary

MSR-VTT

10,000 videos
- 6,512 train
- 498 validation
- 2,990 test

~ 20 captions per video

~ 14K words vocabulary

TRECVID 2020

9,185 videos
- 7,485 train
- 1,700 test

2 ~ 5 captions per video

~ 11K words vocabulary

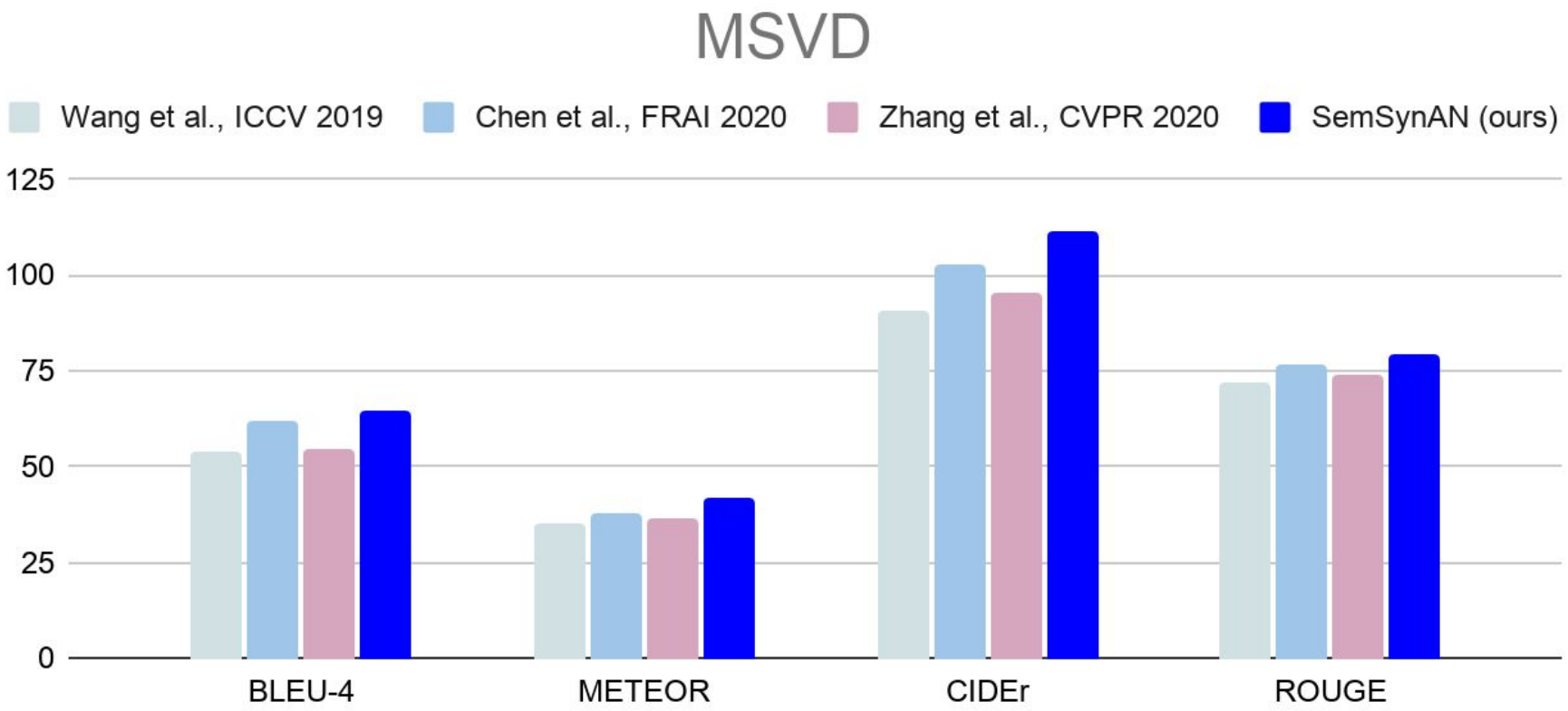
Visual Features:

2D-CNN: ResNet-152 pre-trained on ImageNet

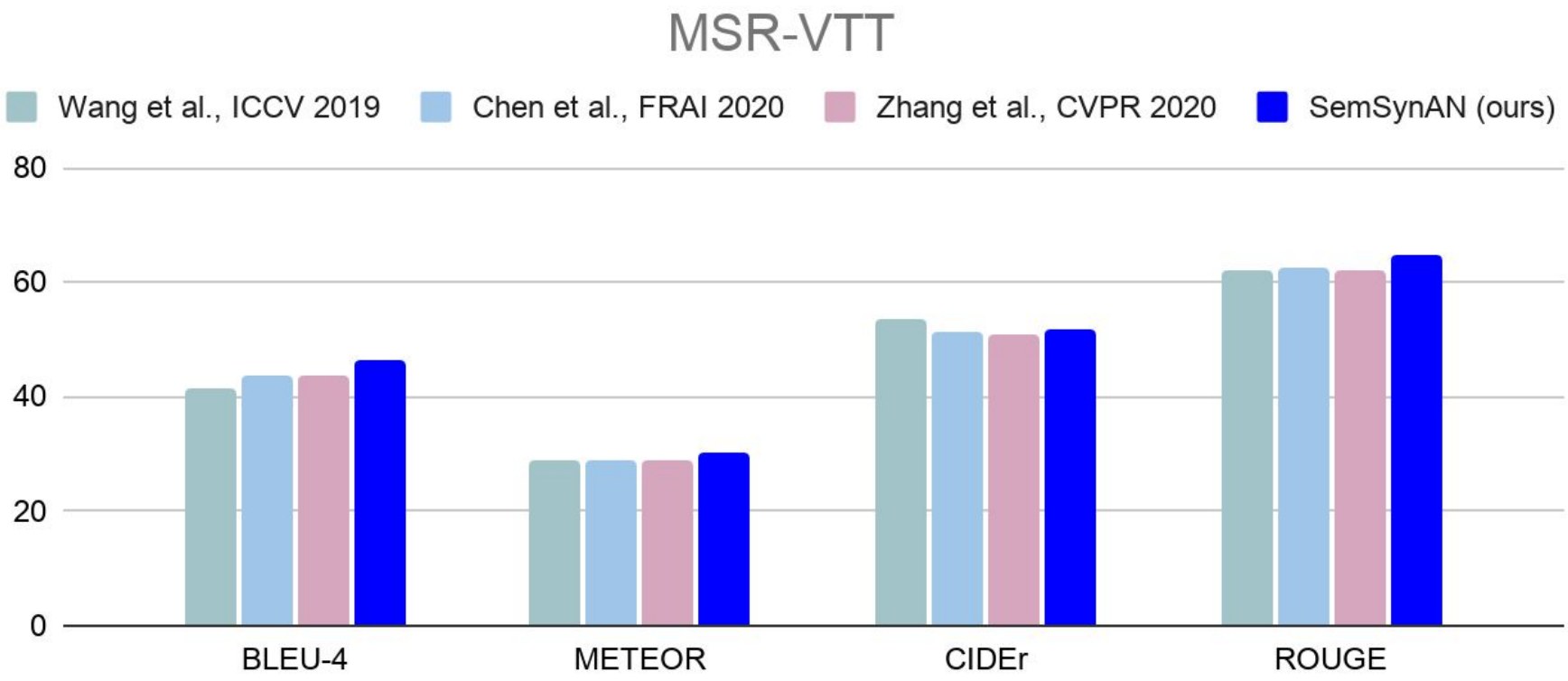
3D-CNN: ECO and R(2+1)D, both pre-trained on Kinetics-400

PyTorch

Results - Comparison with State of the Art



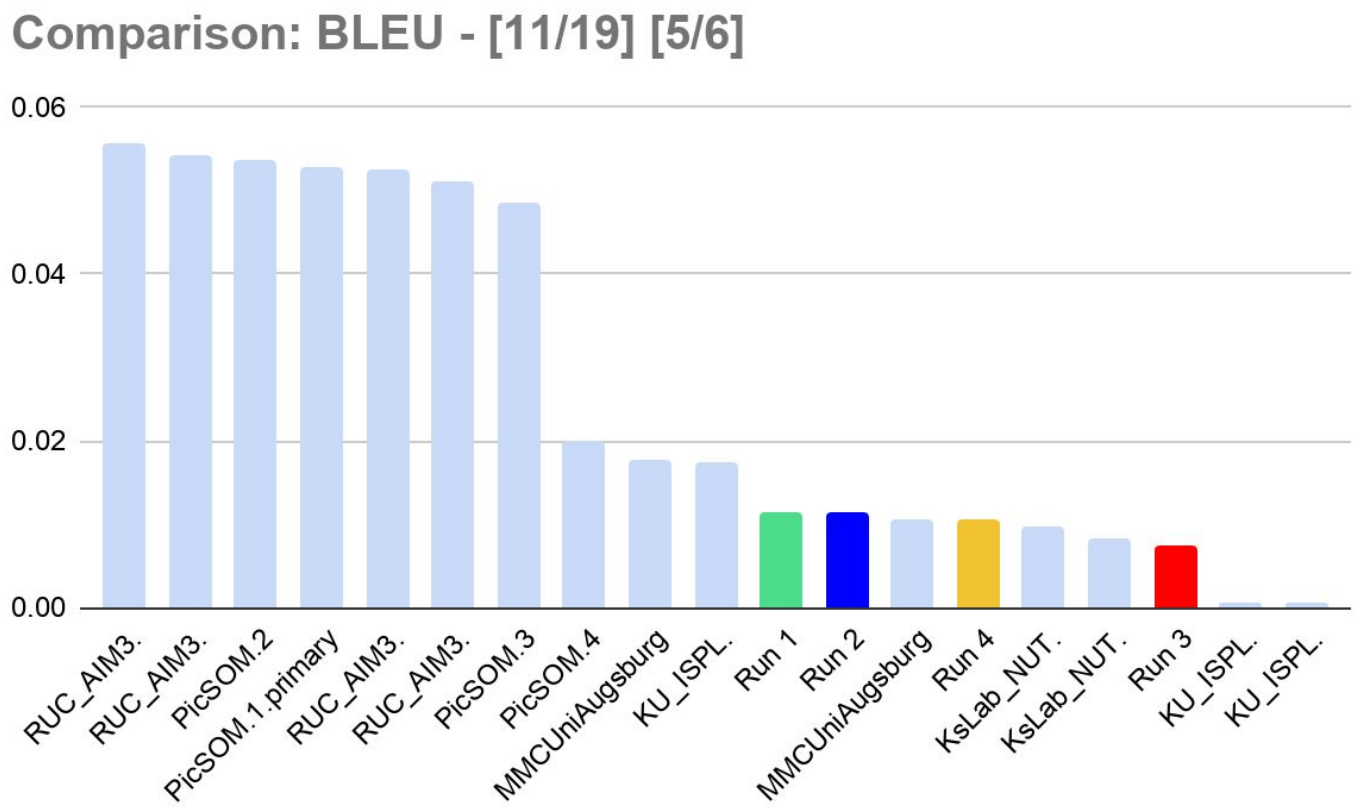
Results - Comparison with State of the Art



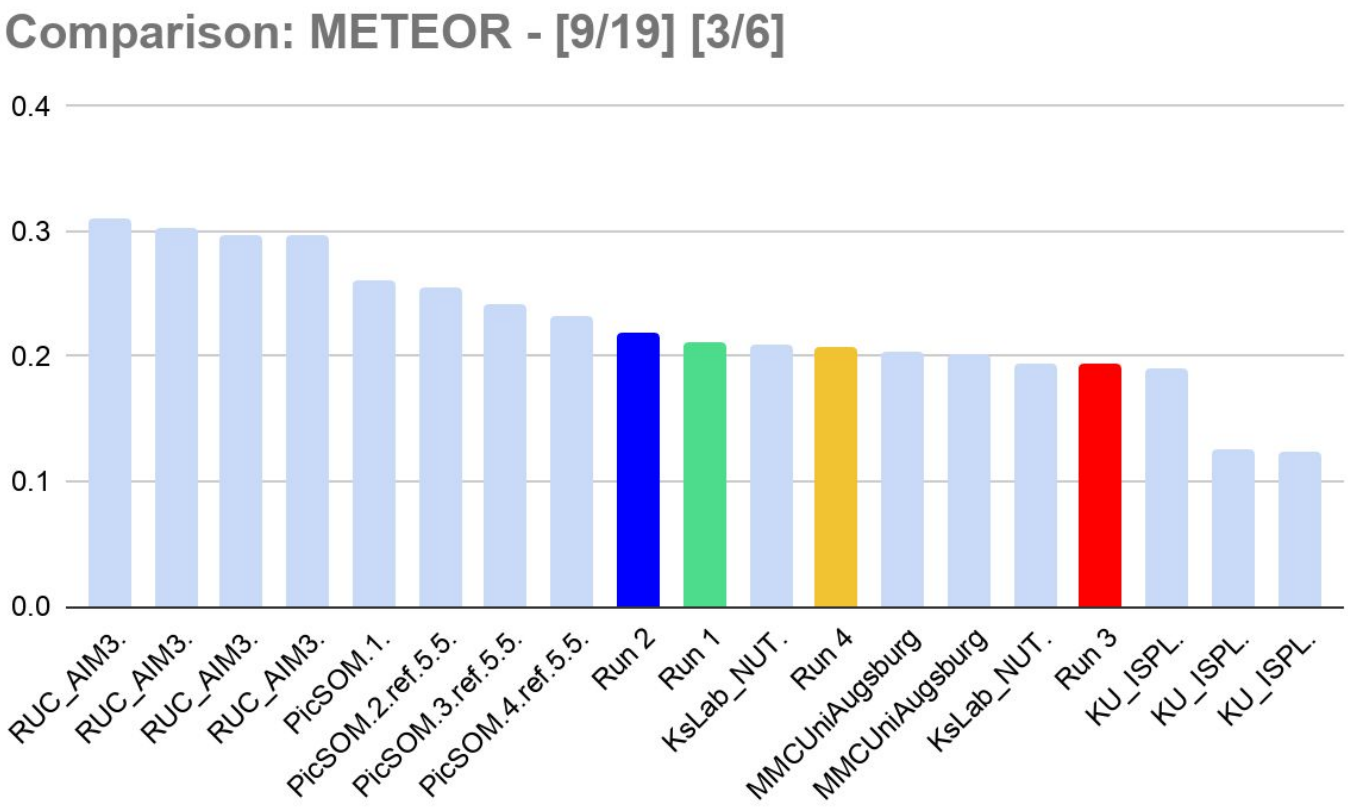
Results - TRECVID 2020

Run	Training Dataset	Validation	epochs	BLEU-4	METEOR	CIDEr	CIDEr-D	SPICE
1	MSRVTT + VTT20 (80%)	VTT20(20%)	40	**0.0115**	0.2105	0.125	0.06	0.057
2	MSRVTT + VTT20	-	29	0.0113	**0.2187**	**0.136**	**0.065**	**0.06**
3	MSRVTT + VTT20 (80%) + VATEX	VTT20(20%)	3	0.0075	0.1938	0.087	0.047	0.04
4	MSRVTT + VTT20	-	46	0.0105	0.2071	0.124	0.062	0.055

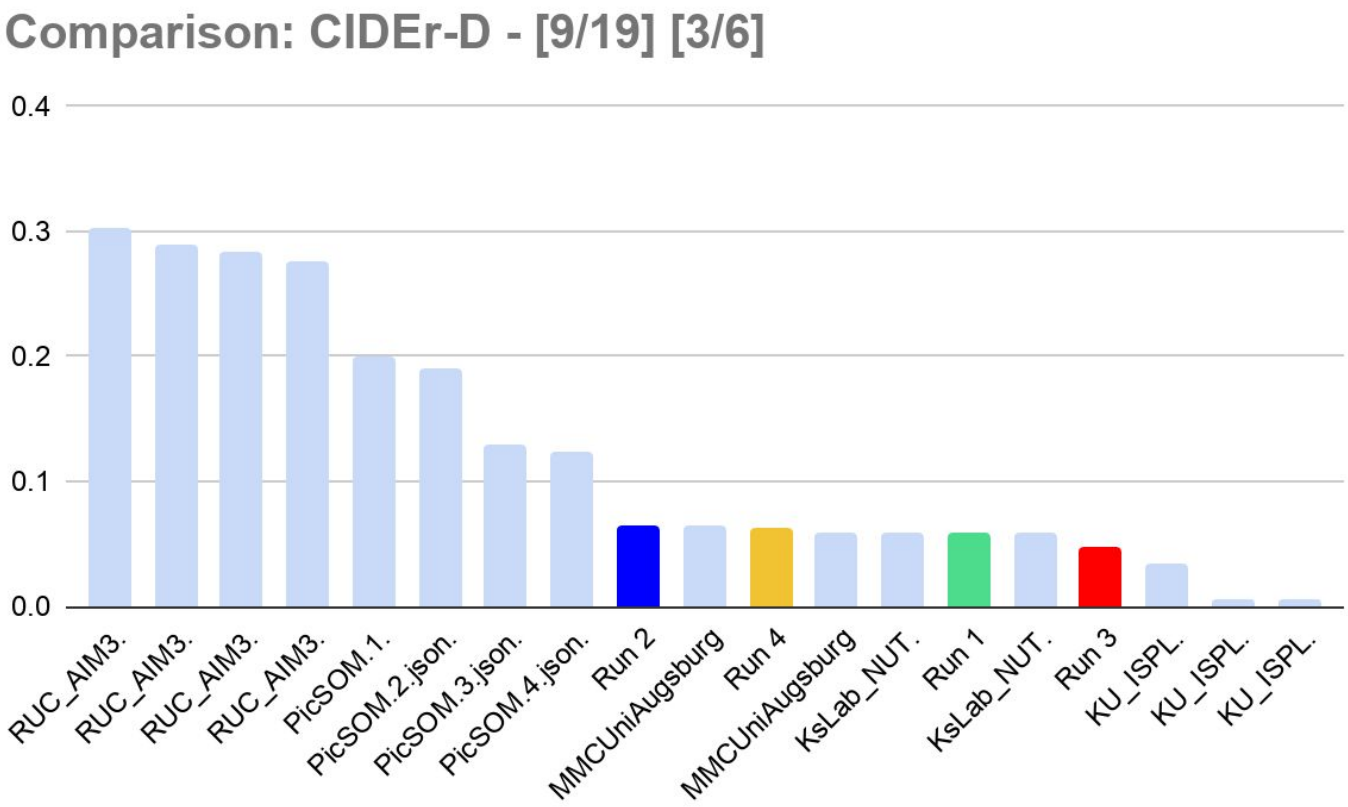
Results - TRECVID 2020



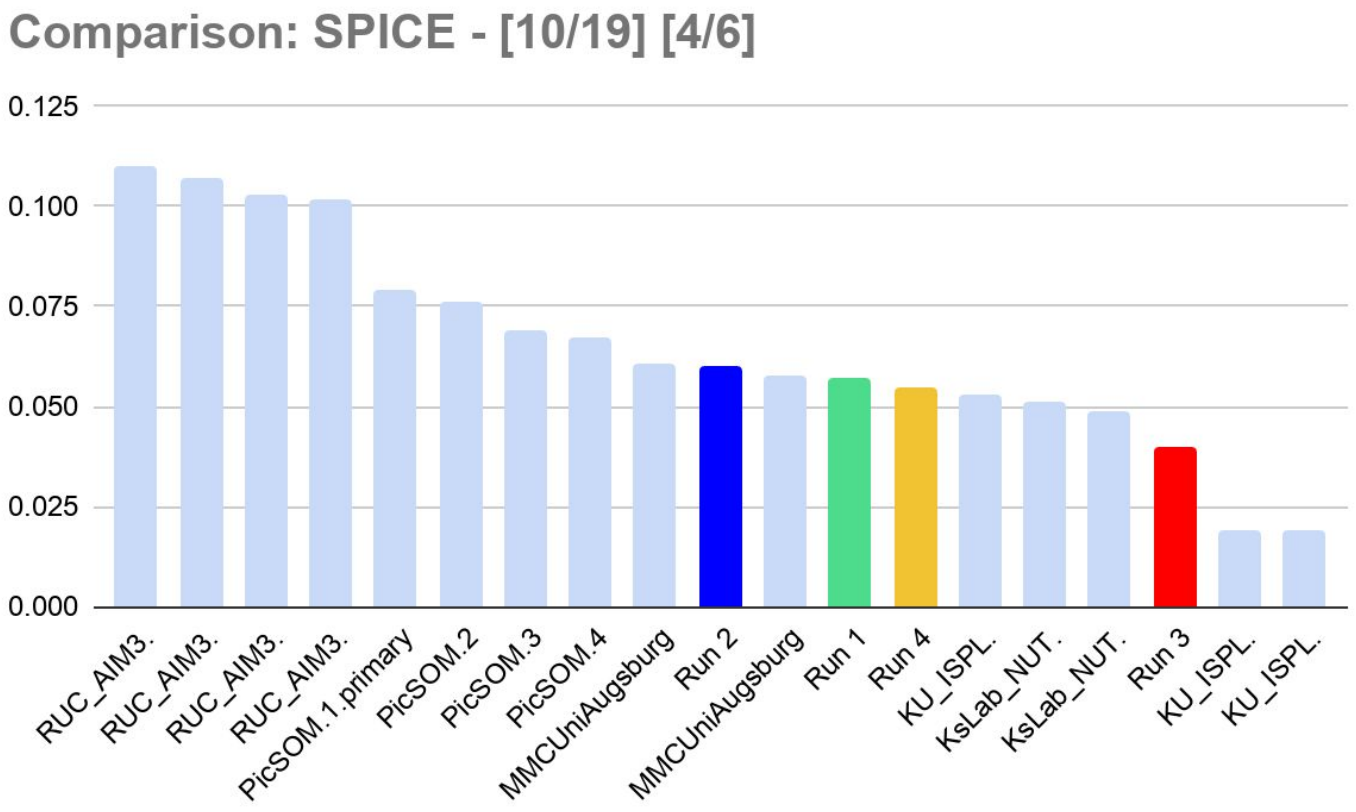
Results - TRECVID 2020



Results - TRECVID 2020



Results - TRECVID 2020



Qualitative Analysis: MSVD

Ours: a woman is applying **makeup on her face**

NN IN PRP$ NN

GT1: a woman is applying **makeup on her face**

GT2: a woman is powdering **her face**

w/o syntactic representation: a woman is applying eye shadow

Qualitative Analysis: MSVD

***Ours*: a man is pouring salsa into a bowl**

NN IN DT NN

***GT1*: a man is putting food on a plate**

***GT2*: the man is pouring salsa over the pasta**

***w/o syntactic representation*: a man is pouring sauce over spaghetti sauce over spaghetti sauce**

Qualitative Analysis: MSVD

***Ours*: a man** and **woman are riding a motorcycle**

DT NN CC NN VBP VBG DT NN

GT1: a man and woman are riding a motorcycle

GT2: a man and a woman are riding a motorcycle

w/o syntactic representation: a man is riding a motorcycle.

Conclusions and Work Plan for TRECVID 2021

1. Paying more attention to syntax improves the quality of descriptions.
2. Controlling the semantic meaning and syntactic structure of generated captions guarantees the contextual relation between the words in the sentence.

As **feature work**, we consider to Improve visual-syntactic embedding by learning to relate syntactic information to a graph-based representation of visual content.

Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding

jeperez@dcc.uchile.cl

@jes_prz

Code/Features/Models available on GitHub
https://github.com/jsprz/visual_syntactic_embedding_video_captioning

Millennium Institute
Foundational
Research on Data