University Defence Research Collaboration (UDRC)
Signal Processing in the Information Age

Multimodal Fusion:
Natural Language and Vision

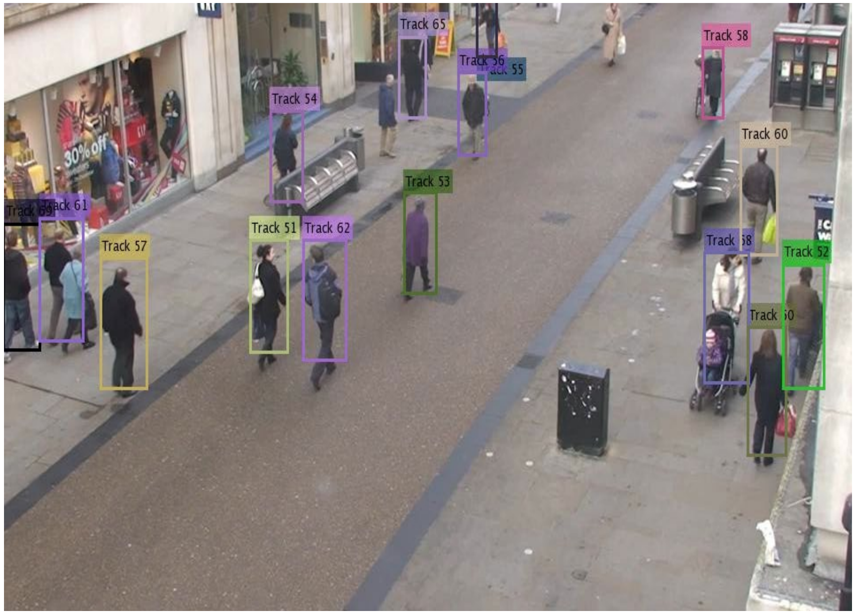
Academic Lead: Prof. Neil Robertson
Research Fellow: Dr Shiyang Yan

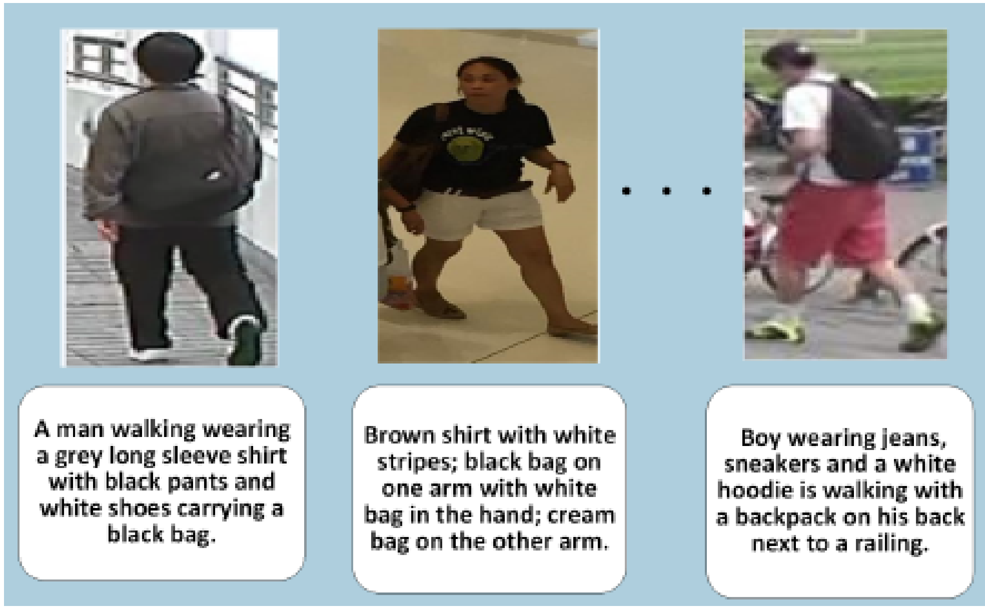
PhDs: Rachael Abbott, Jian Gao
Academics: Tim Hospedales, Joao Mota, Sotos Tsaftaris

[dstl]

EPSRC
Engineering and Physical Sciences
Research Council

Motivations for Defence Applications





A man walking wearing a grey long sleeve shirt with black pants and white shoes carrying a black bag.

Brown shirt with white stripes; black bag on one arm with white bag in the hand; cream bag on the other arm.

Boy wearing jeans, sneakers and a white hoodie is walking with a backpack on his back next to a railing.

- In defence or suivallance domain, multimodal data help to comprehensively and accurately detect **potential danger**. This is **advantageous** than single modal data which only provide **one aspect** of the situation.

- However, data and annotations are expensive. Hence, we provide general solutions to **generate** text data based on visual data, which are then used for various defence applications.

We start with multi-modal data

Multi-modal scene understanding

- Semi-supervised learning - few labels - transfer knowledge
- Multi-task: Classify and Segment - should reinforce each other
- Examples from one domain, or only one modality
    - How do we learn the most effective mapping to other domain/modality?
    - How do we assign and use semantic labels to help?

Specifically, these works have been done in the WP3.3 so far:

- **Image captioning:**

  SC-RANK: Improving Convolutional Image Captioning with Self-Critical Learning and Ranking Metric-based Reward--The application of reinforcement learning in multi-modal data
- **Visual Paragraph Generation:**

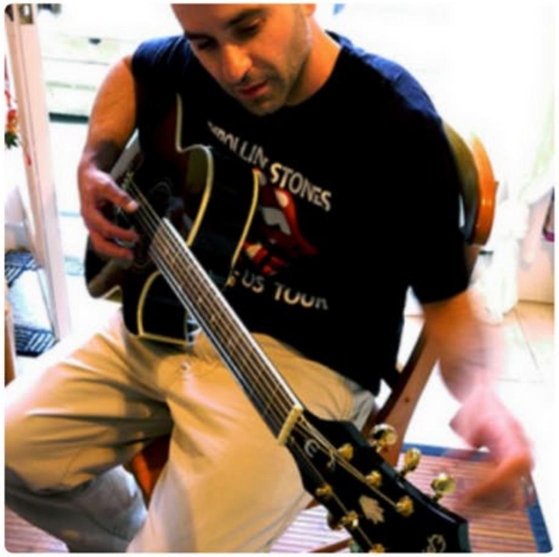
  ParaCNN: Visual Paragraph generation via adversarial twin Contextual CNNs--Try to solve the long sequence problem in multi-modal domain
- **Multimodal and Multitask Learning for Person Re-ID:**

  We leverage the description data to aid the visual person Re-ID and propose a multitask learning framework to train the person Re-ID and image captioning at the same time.

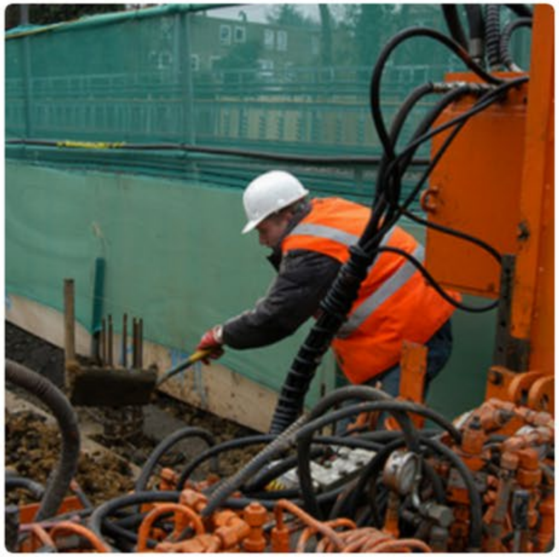
SC-RANK (Image Captioning)

Image captioning is a popular research topic in both computer vision and natural language processing. It has attracted much attention from researchers in recent years.

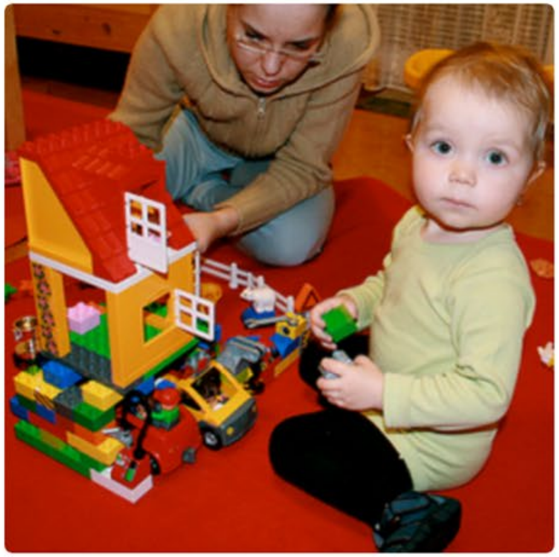
We can see some examples on how the current algorithm can generate natural language descriptions given an image.



"man in black shirt is playing guitar."

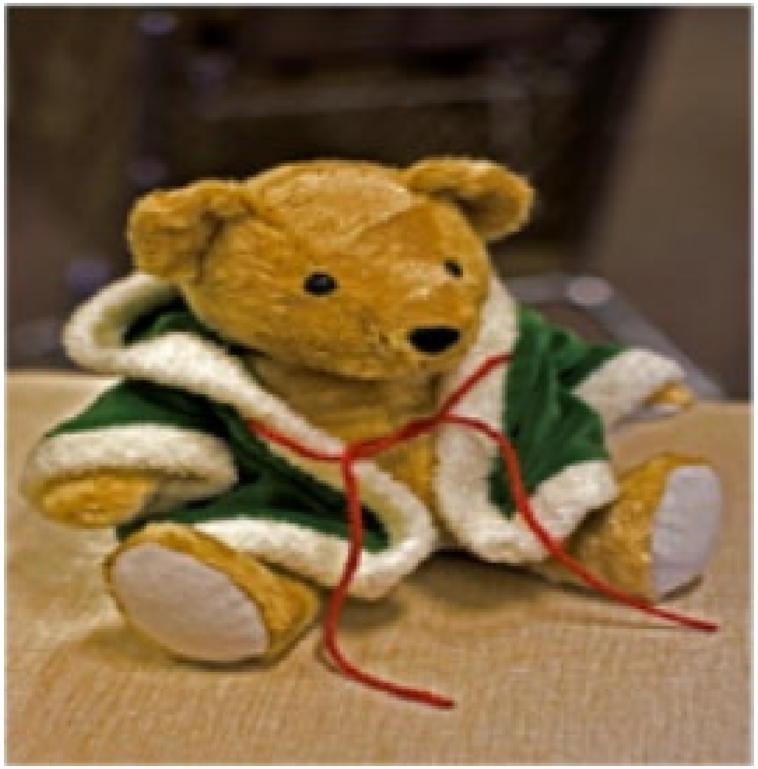


"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."

SC-RANK (Image Captioning)

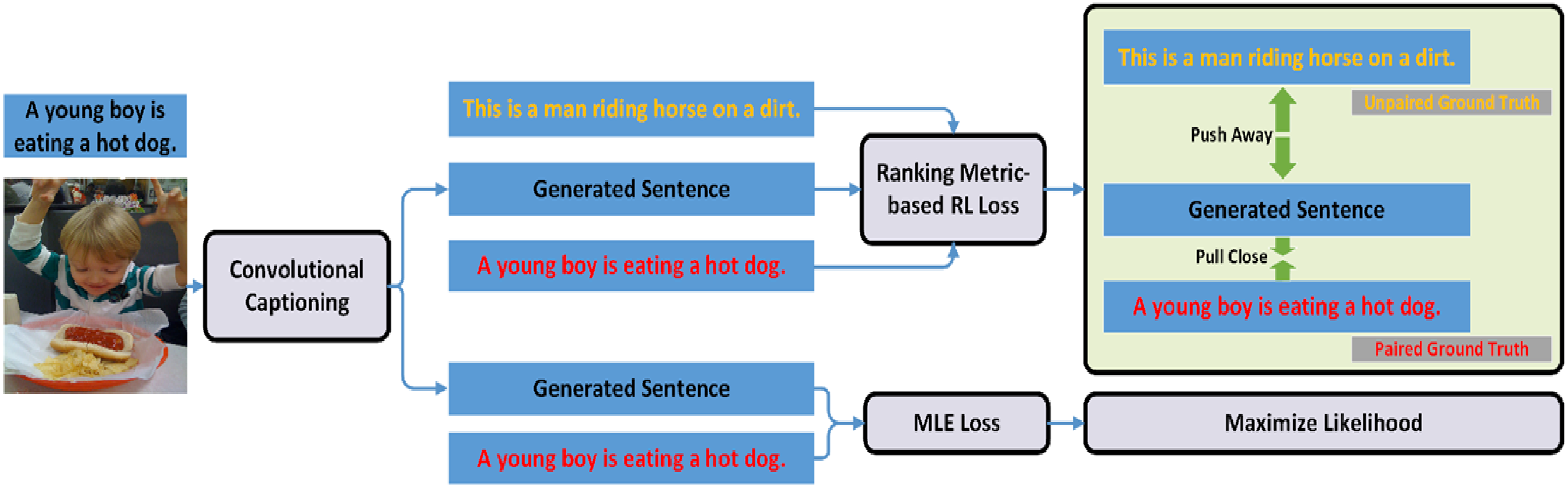


MLE: A stuffed teddy bear sitting on a chair.

Ours: A teddy bear is wearing a bow tie sitting in a bed.

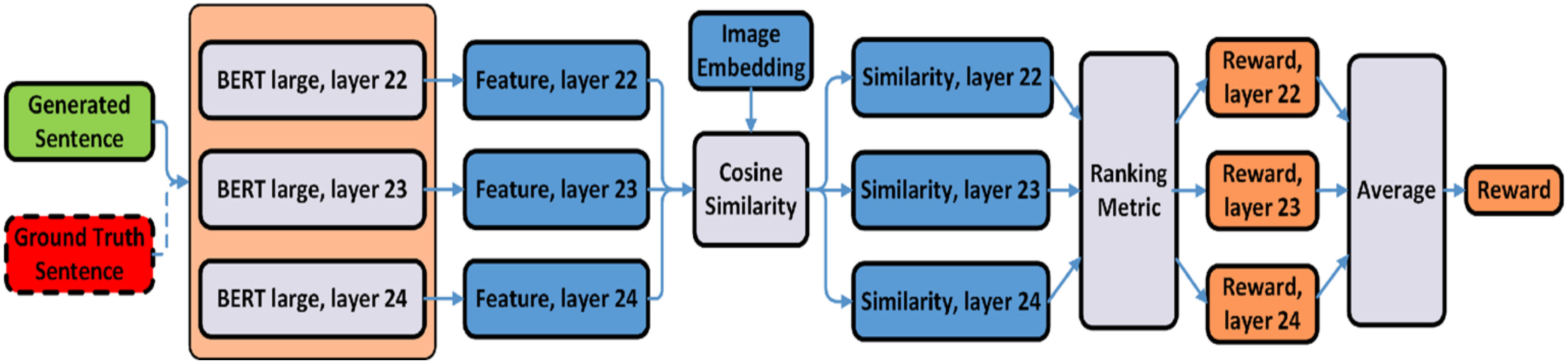
SC-RANK (Image Captioning)

- A novel ranking **metric-based reward**, denoted as **SC-RANK,** is proposed to achieve diversified sentence generation. The experimental results show improvement by using the proposed SC-RANK.


A young boy is eating a hot dog.
Convolutional Captioning
This is a man riding horse on a dirt.
Generated Sentence
A young boy is eating a hot dog.
Ranking Metric-based RL Loss
This is a man riding horse on a dirt.
Unpaired Ground Truth
Push Away
Generated Sentence
Pull Close
A young boy is eating a hot dog.
Paired Ground Truth
Generated Sentence
A young boy is eating a hot dog.
MLE Loss
Maximize Likelihood

SC-RANK (Image Captioning)

- The sentence embedding is from a **BERT,** and the features from several layers from the BERT are used to compute the reward.

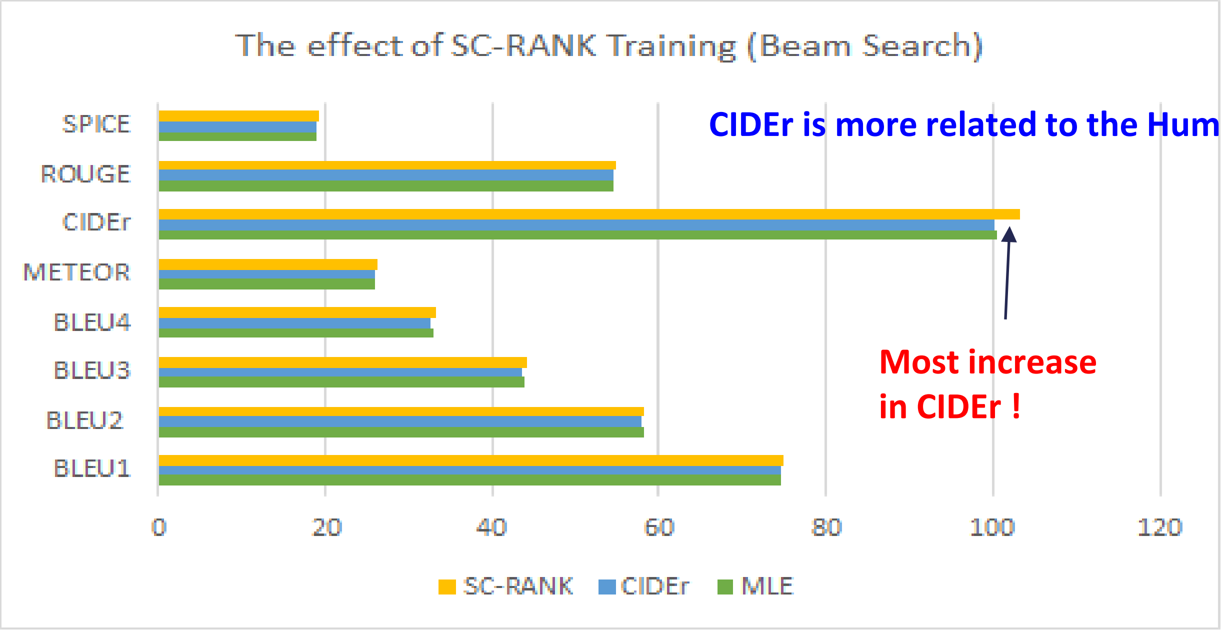


SC-RANK (Image Captioning)

- Specifically, the MLE indicates the baseline, CIDEr indicates the conventional self-critical learning. SC-RANK is the proposed method.

- An interesting phenomenon is that unlike conventional self-critical learning, our SC-RANK does not prevent beam search from further raising the performance.

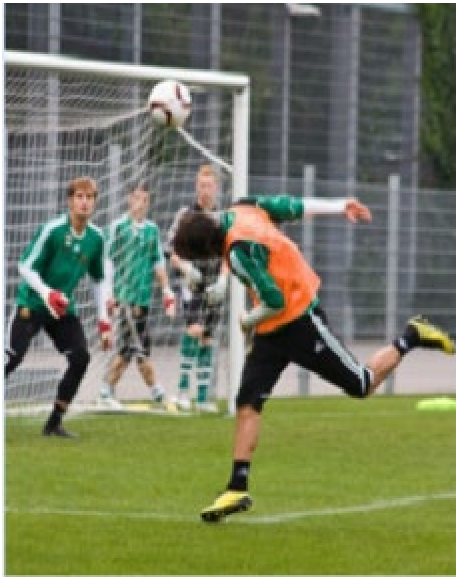
- As shown in the graph, our SC-RANK improves both the MLE baseline and the conventional self-critical learning in scenarios of greedy decoding and beam search, which proves the effectiveness of the SC-RANK.


The effect of SC-RANK Training (Beam Search)

CIDEr is more related to the Human Consensus!

Most increase in CIDEr !

SC-RANK (Image Captioning)



MLE Baseline:
A soccer player jumps to catch a soccer ball.

Ours 1 (BERT):
Several men playing soccer in a soccer field.

Ours 2 (InferSent):
A soccer player running to catch a soccer ball.

ParaCNN (Image Paragraph generation)

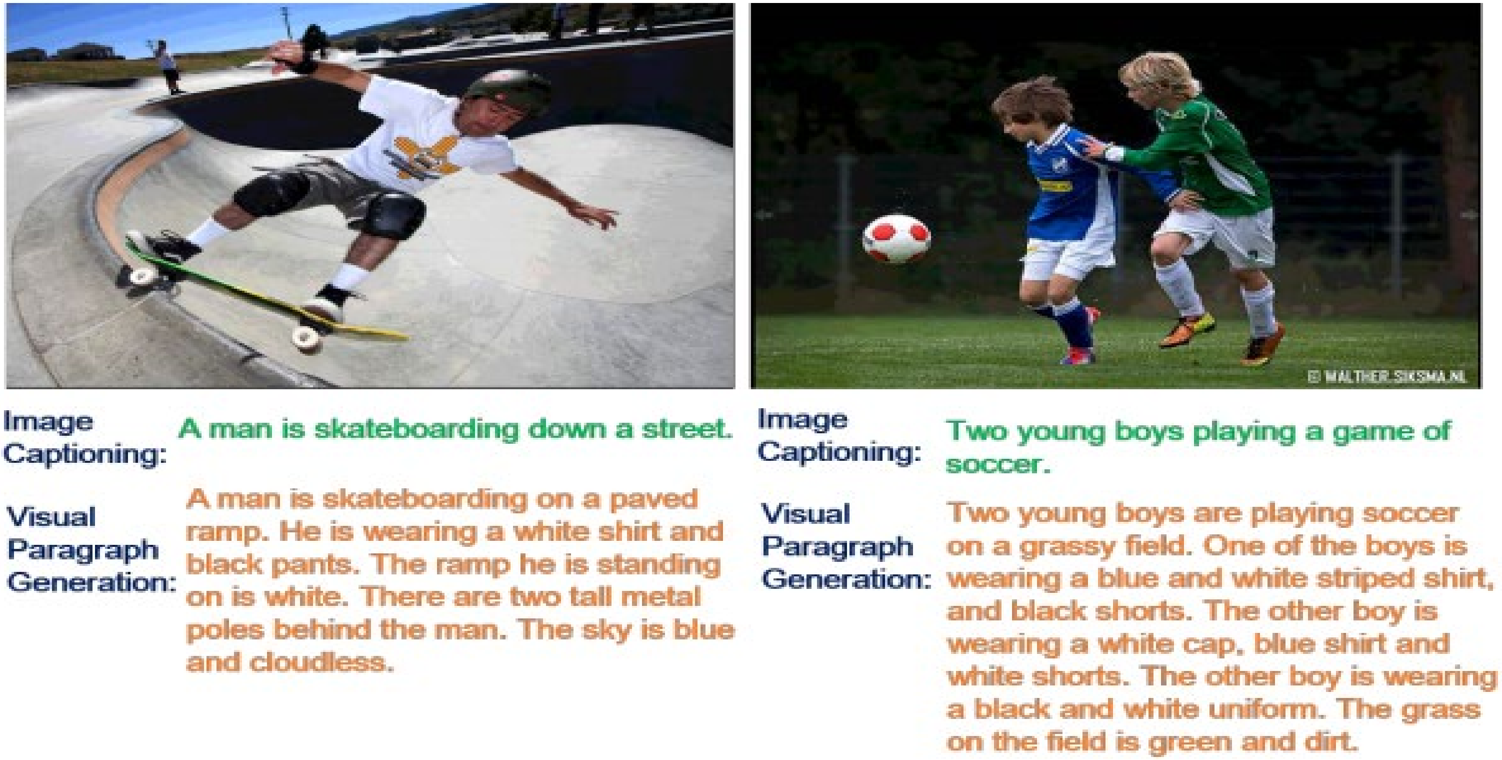


Image Captioning: A man is skateboarding down a street.

Visual Paragraph Generation: A man is skateboarding on a paved ramp. He is wearing a white shirt and black pants. The ramp he is standing on is white. There are two tall metal poles behind the man. The sky is blue and cloudless.

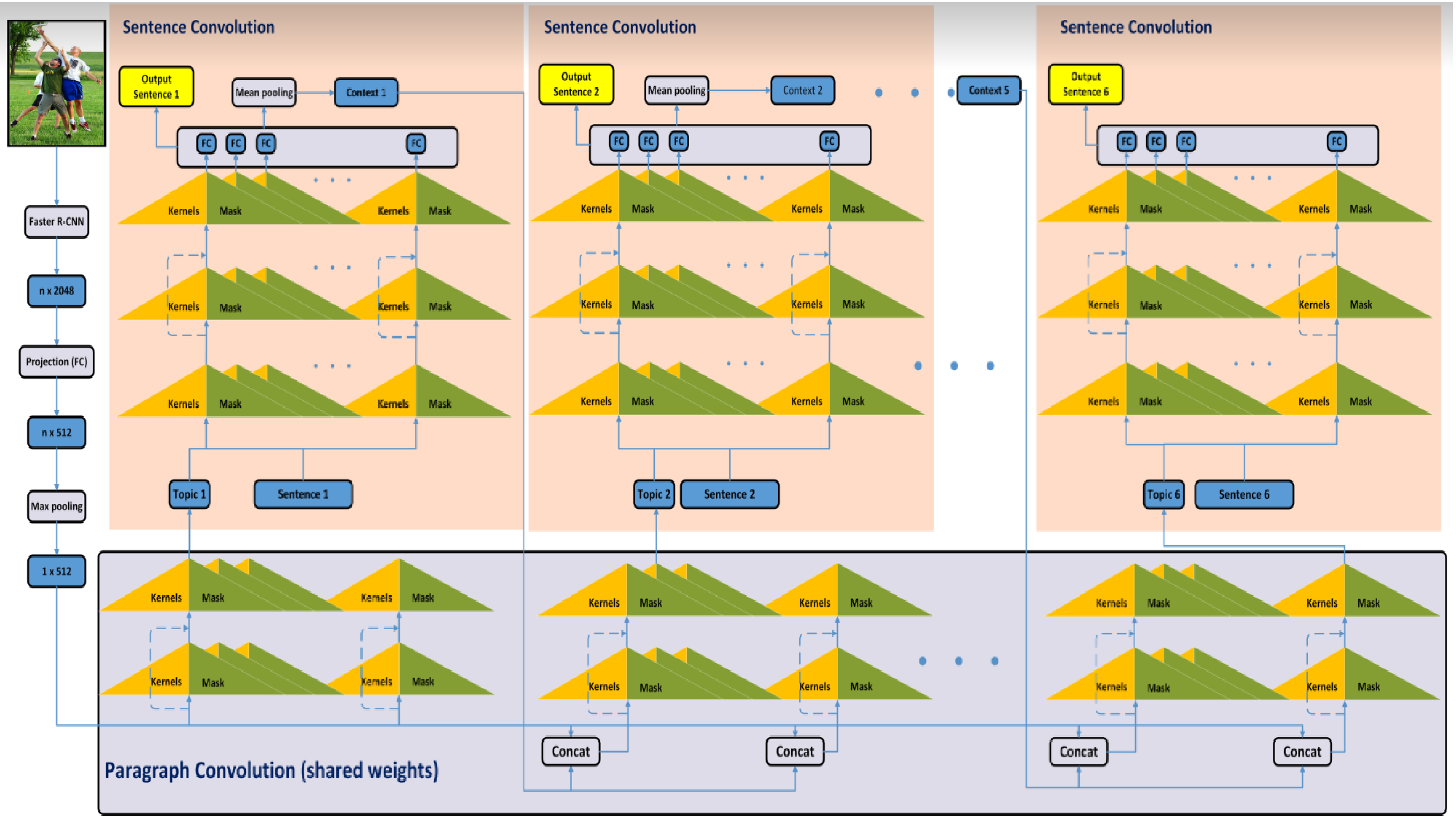
Image Captioning: Two young boys playing a game of soccer.

Visual Paragraph Generation: Two young boys are playing soccer on a grassy field. One of the boys is wearing a blue and white striped shirt, and black shorts. The other boy is wearing a white cap, blue shirt and white shorts. The other boy is wearing a black and white uniform. The grass on the field is green and dirt.

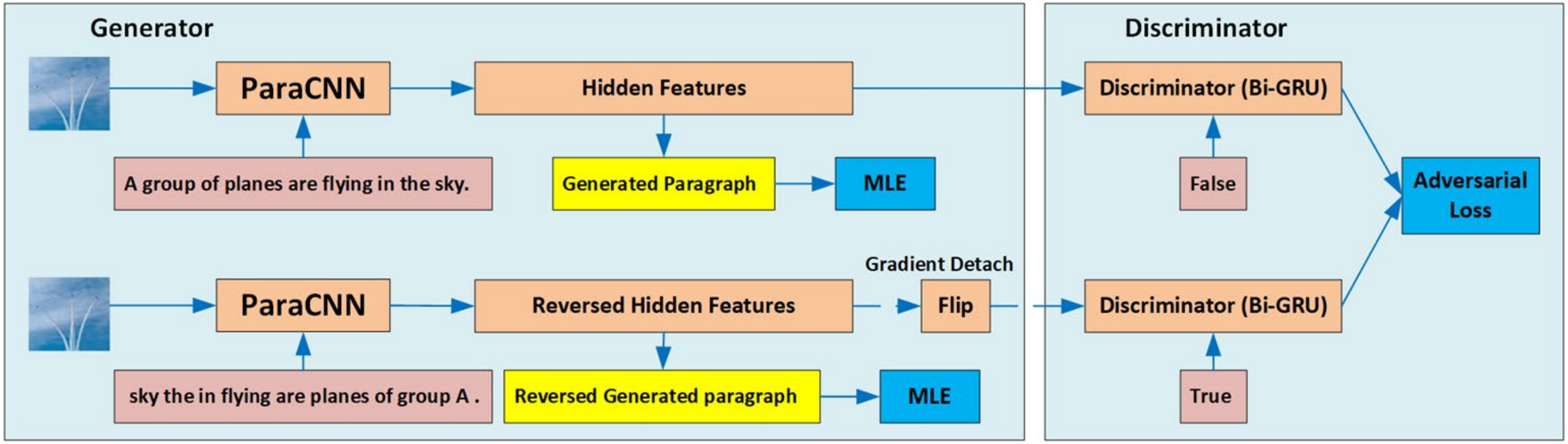
- Image captioning is a well-developed problem but a single short sentence with **less than 20 words** is not enough to describe the full content of an image which could be very informative. This is especially true for **defence**, which requires **high accuracy**.

- We propose a **novel CNN architecture** to effectively and efficiently solve the problem of long sequence modelling, **without using any RNN-like** structure at all, which is more time efficient and effective.

ParaCNN (Image Paragraph generation)



ParaCNN (Image Paragraph generation)

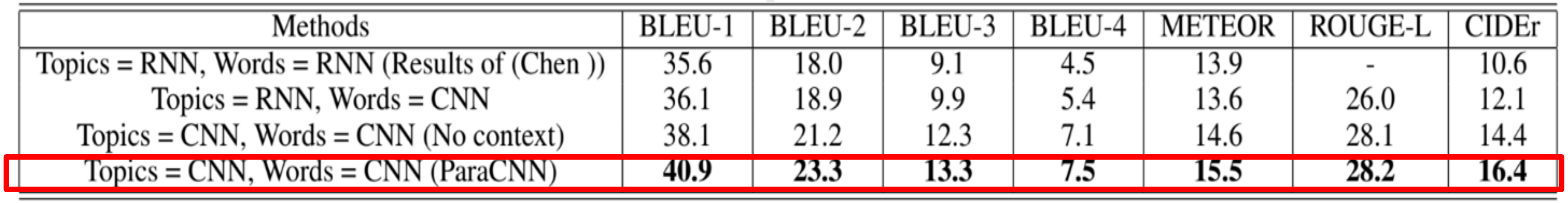


- In addition, We propose an **adversarial twin net training** scheme, to make the distributions of the forwarding network close to the backwarding network.

- The one used in inference is forwarding networks whilst the backwarding network’s knowledge is transferred to the forwarding one during training stage.

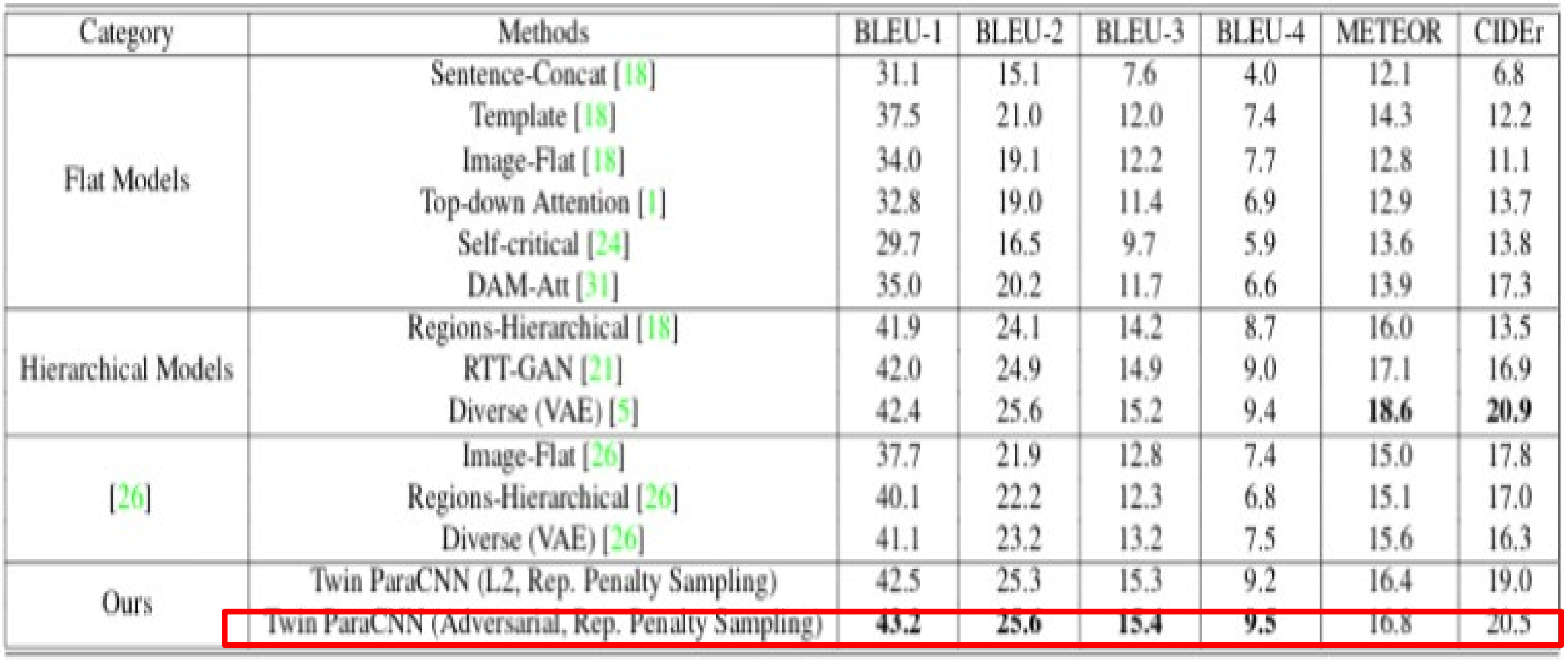
ParaCNN (Image Paragraph generation)

Table 1: Ablation Studies on the different schemes: our ParaCNN with contextual information yields the best performance.



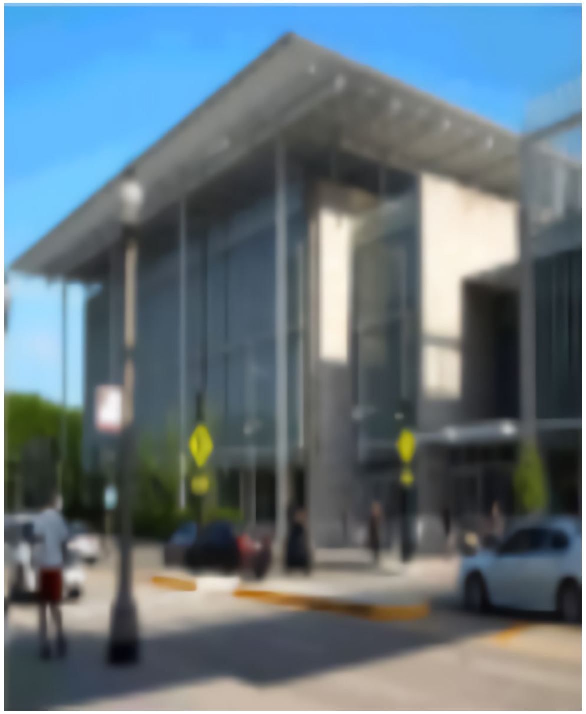
Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Topics = RNN, Words = RNN (Results of (Chen))	35.6	18.0	9.1	4.5	13.9	-	10.6
Topics = RNN, Words = CNN	36.1	18.9	9.9	5.4	13.6	26.0	12.1
Topics = CNN, Words = CNN (No context)	38.1	21.2	12.3	7.1	14.6	28.1	14.4
Topics = CNN, Words = CNN (ParaCNN)	**40.9**	**23.3**	**13.3**	**7.5**	**15.5**	**28.2**	**16.4**

Table 2: State-of-the-art performance has been achieved.



Category	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Flat Models	Sentence-Concat [18]	31.1	15.1	7.6	4.0	12.1	6.8
	Template [18]	37.5	21.0	12.0	7.4	14.3	12.2
	Image-Flat [18]	34.0	19.1	12.2	7.7	12.8	11.1
	Top-down Attention [1]	32.8	19.0	11.4	6.9	12.9	13.7
	Self-critical [24]	29.7	16.5	9.7	5.9	13.6	13.8
	DAM-Att [31]	35.0	20.2	11.7	6.6	13.9	17.3
Hierarchical Models	Regions-Hierarchical [18]	41.9	24.1	14.2	8.7	16.0	13.5
	RTT-GAN [21]	42.0	24.9	14.9	9.0	17.1	16.9
	Diverse (VAE) [5]	42.4	25.6	15.2	9.4	**18.6**	**20.9**
[26]	Image-Flat [26]	37.7	21.9	12.8	7.4	15.0	17.8
	Regions-Hierarchical [26]	40.1	22.2	12.3	6.8	15.1	17.0
	Diverse (VAE) [26]	41.1	23.2	13.2	7.5	15.6	16.3
Ours	Twin ParaCNN (L2, Rep. Penalty Sampling)	42.5	25.3	15.3	9.2	16.4	19.0
	Twin ParaCNN (Adversarial, Rep. Penalty Sampling)	**43.2**	**25.6**	**15.4**	**9.5**	16.8	20.5

ParaCNN (Image Paragraph Generation)



Ground Truth:
A large building with bars on the windows in front of it. There is people walking in front of the building. There is a street in front of the building with many cars on it.

Ours:
The picture is taken outside on a sunny day. A large building can be seen along a sidewalk. Vehicles can be seen parked on a road near the sidewalk. Two vehicles can be seen driving on the road. Tall green trees are standing on the side of the road.

The red text is novel concepts even neglected in the ground-truth annotations.

An application of multimodal fusion in defence ---- Person Re-identification

- Person Re-identification is a visual surveillance task which is very important in defence and security.

(a)

(b)

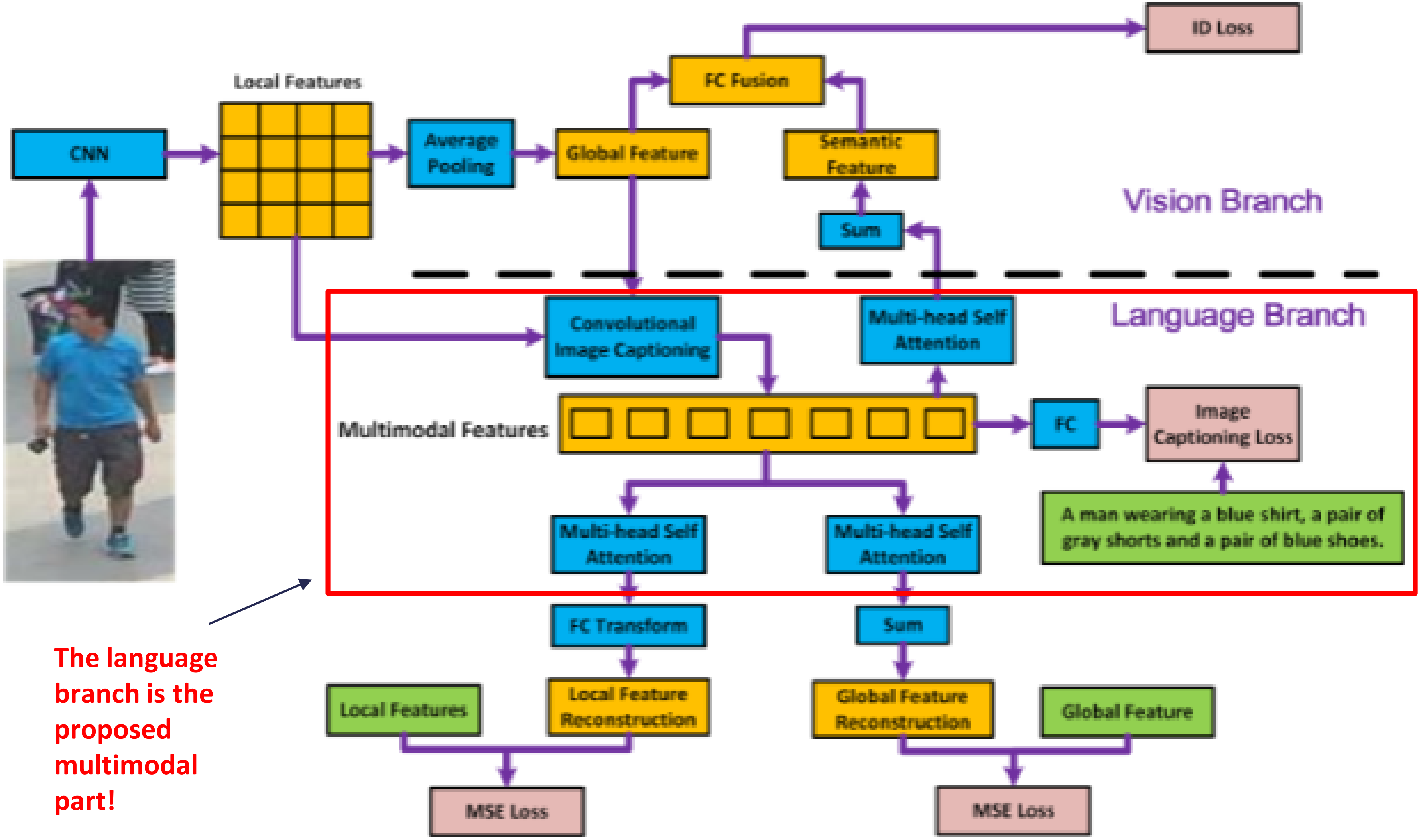
Multitask multimodal Fusion for Person Re-ID

- We propose to use **multimodal data** with an extra modality of **natural language description** of a person to perform person Re-ID.

- We **do not need language description** of a person image during testing which is an appealing property.

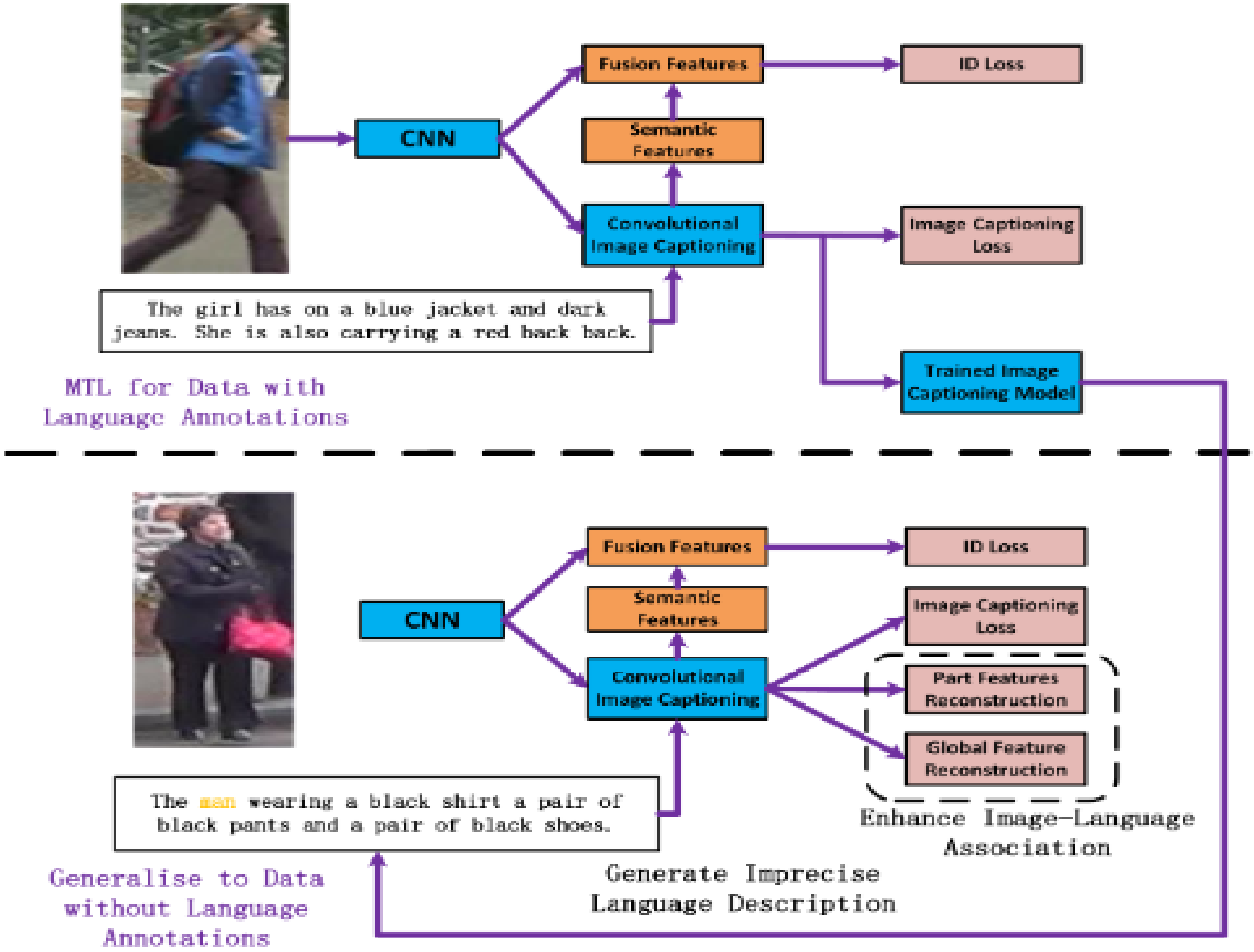
- Even for the training data that lacks language description, we can use pre-trained image captioning model to **generate the descriptions.** But these descriptions are usually **noisy.**

Multitask multimodal Fusion for Person Re-ID


ID Loss
FC Fusion
Local Features
CNN
Average Pooling
Global Feature
Semantic Feature
Sum
Vision Branch
Language Branch
Convolutional Image Captioning
Multi-head Self Attention
Multimodal Features
FC
Image Captioning Loss
A man wearing a blue shirt, a pair of gray shorts and a pair of blue shoes.
Multi-head Self Attention
Multi-head Self Attention
FC Transform
Sum
Local Features
Local Feature Reconstruction
Global Feature Reconstruction
Global Feature
MSE Loss
MSE Loss

The language branch is the proposed multimodal part!

Multitask multimodal Fusion for Person Re-ID



CNN

Fusion Features

Semantic Features

Convolutional Image Captioning

ID Loss

Image Captioning Loss

Trained Image Captioning Model

The girl has on a blue jacket and dark jeans. She is also carrying a red back back.

MTL for Data with Language Annotations

CNN

Fusion Features

Semantic Features

Convolutional Image Captioning

ID Loss

Image Captioning Loss

Part Features Reconstruction

Global Feature Reconstruction

Enhance Image-Language Association

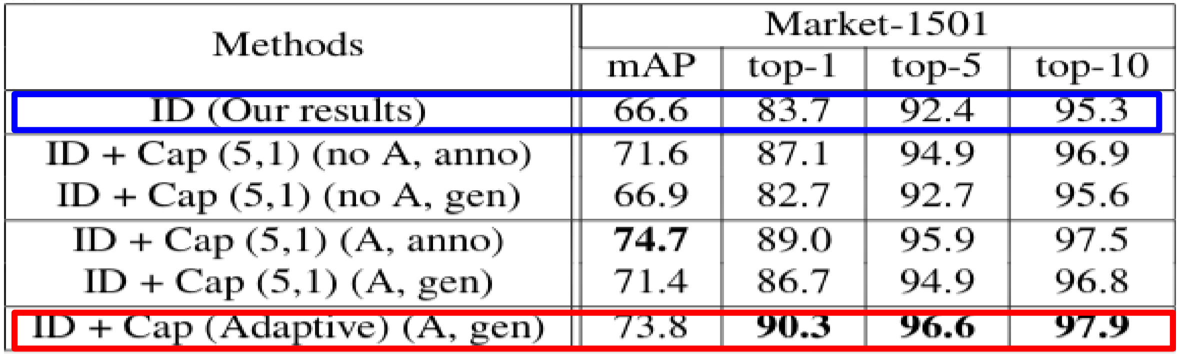
The man wearing a black shirt a pair of black pants and a pair of black shoes.

Generate Imprecise Language Description

Generalise to Data without Language Annotations

Multitask multimodal Fusion for Person Re-ID

Table 1. Ablation Studies on Market-1501 with Language Annotations



Methods	Market-1501			
	mAP	top-1	top-5	top-10
ID (Our results)	66.6	83.7	92.4	95.3
ID + Cap (5,1) (no A, anno)	71.6	87.1	94.9	96.9
ID + Cap (5,1) (no A, gen)	66.9	82.7	92.7	95.6
ID + Cap (5,1) (A, anno)	**74.7**	89.0	95.9	97.5
ID + Cap (5,1) (A, gen)	71.4	86.7	94.9	96.8
ID + Cap (Adaptive) (A, gen)	73.8	**90.3**	**96.6**	**97.9**

Our proposed model increase the baseline method with roughly 7 percent on the mAP result, with 7 percent on the ranking-1 result.

Annotations:
The girl is wearing a yellow blouse with black pants. She is riding a bicycle.

Sampled during testing:
woman wearing a white short sleeved shirt and black pants is riding a bike.

Annotations:
the boy walking on concrete is wearing a purple shirt opened at the neck and blue jeans.

Sampled during testing:
man is wearing a purple shirt and black is carrying a black backpack over his left.

Our method can sample new captioning, thus generate semantic features to facilitate the re-ID task, during the test time.

Multitask multimodal Fusion for Person Re-ID

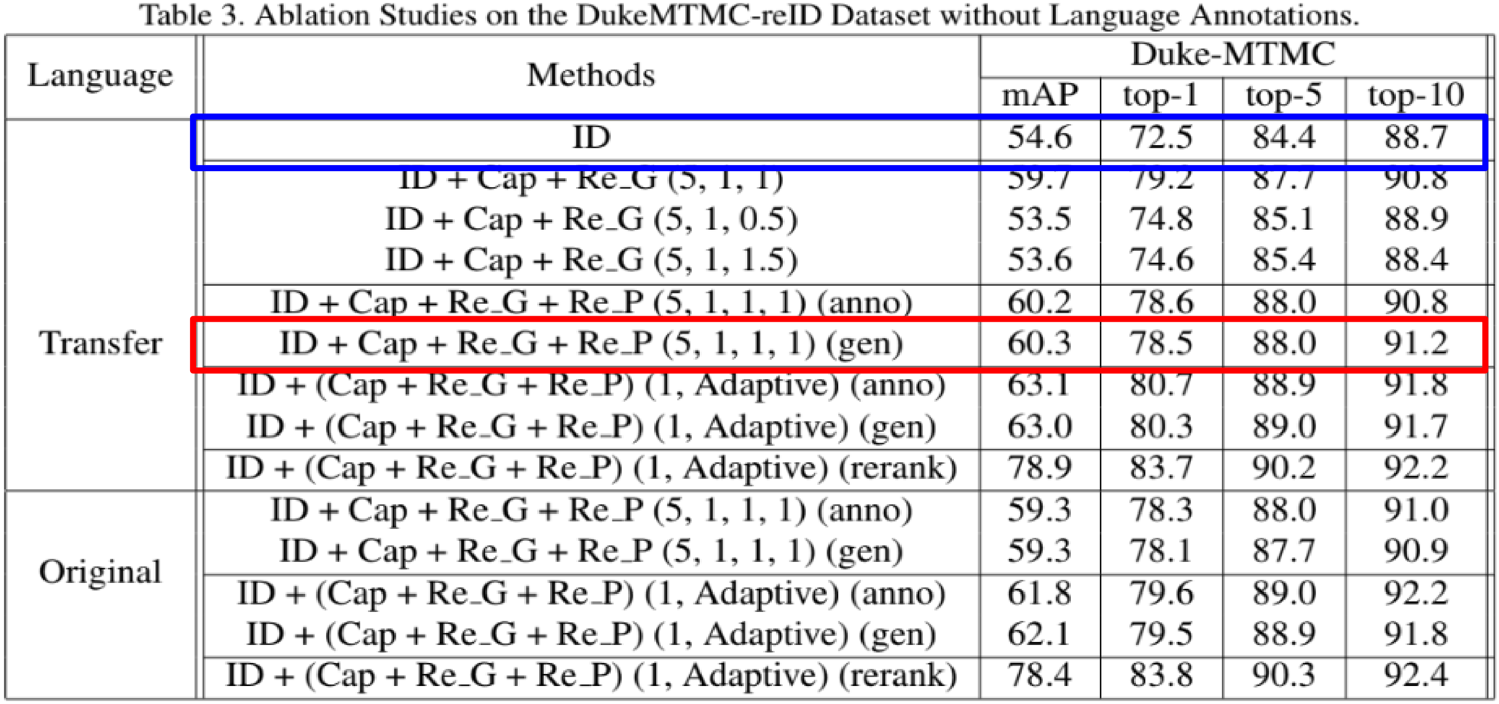
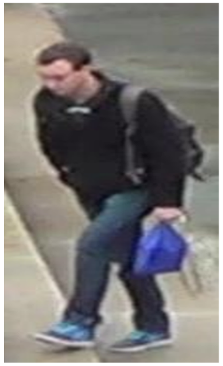


Table 3. Ablation Studies on the DukeMTMC-reID Dataset without Language Annotations.

Language	Methods	Duke-MTMC			
		mAP	top-1	top-5	top-10
Transfer	ID	54.6	72.5	84.4	88.7
	ID + Cap + Re_G (5, 1, 1)	59.7	79.2	87.7	90.8
	ID + Cap + Re_G (5, 1, 0.5)	53.5	74.8	85.1	88.9
	ID + Cap + Re_G (5, 1, 1.5)	53.6	74.6	85.4	88.4
	ID + Cap + Re_G + Re_P (5, 1, 1, 1) (anno)	60.2	78.6	88.0	90.8
	ID + Cap + Re_G + Re_P (5, 1, 1, 1) (gen)	60.3	78.5	88.0	91.2
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (anno)	63.1	80.7	88.9	91.8
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (gen)	63.0	80.3	89.0	91.7
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (rerank)	78.9	83.7	90.2	92.2
Original	ID + Cap + Re_G + Re_P (5, 1, 1, 1) (anno)	59.3	78.3	88.0	91.0
	ID + Cap + Re_G + Re_P (5, 1, 1, 1) (gen)	59.3	78.1	87.7	90.9
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (anno)	61.8	79.6	89.0	92.2
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (gen)	62.1	79.5	88.9	91.8
	ID + (Cap + Re_G + Re_P) (1, Adaptive) (rerank)	78.4	83.8	90.3	92.4

For DukeMTMC-reID dataset, we do not have language annotations. We increase the final mAP result by almost 6 percent, and increase the ranking-1 results by 6 percent.

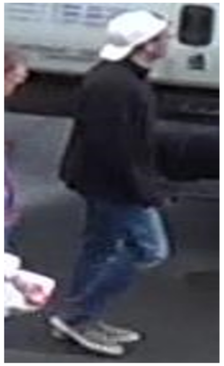
Multitask multimodal Fusion for Person Re-ID



Annotations:
the man is wearing a black shirt a pair of black pants and a pair of black shoes.

Generated:
man is wearing a white shirt a pair of black pants and a pair of black shoes.

Generated via reconstruction:
man is wearing a black shirt a pair of black pants and a pair of black shoes.



Annotations:
the man is wearing a white shirt a pair of blue jeans and a pair of black shoes.

Generated:
man is wearing a black shirt a black pant and a pair of black shoes.

Generated via reconstruction:
man is wearing a white shirt and a pair of black pants and a pair of black shoes

For the dataset such as **DukeMTMC-reID** which **lacks the annotations**, we use the generated text as annotations and use reconstruction loss to compensate the noisiness. The **yellow text** is the wrong concept in the **initial annotations** and red text is the wrong concept happened during the generating process.

Conclusions

- We investigate real-world problems of vision and language, which includes:

1. image captioning, which is to generate a caption based on an image.
2. Image paragraph generation, which is more challenging.
3. We utilize the multimodal information of vision and text descriptions for person re-ID.

- The future works include the multitask learning theory on multimodal fusion and its applications on computer vision problems in the wild.