PicSOM Team’s Lessons Learned in TRECVID 2020 VTT

Jorma Laaksonen and Zixin Guo
Aalto University
Espoo, Finland
2020-12-08

Overview

- Experiments with datasets and features
- Stacked attention captioning model
- Submissions and results
- Observations & issues

Experiments with datasets and features

- We experimented with a selection training datasets ...
  - COCO – actually only images, but we used “fake” video features
  - TGIF
  - VATEX – new addition this year and we wanted to study its advantage
  - (MSR-VTT and MSVD had been used in earlier years, but dropped now)
- … and features
  - ResNet-152
  - ResNext-101
  - I3D
  - C3D
  - (ResNet-101, semantic category, audio and multimodal features had been used earlier, but dropped now)

Comparison of the datasets

dataset	type	items	captions
COCO	images	82783	414113
TGIF	videos	125713	125713
VATEX	videos	41250	825000

“Fake I3D features” for COCO images

- I3D features can be extracted only from videos
- Average of I3D features of the TGIF videos were used as “fake I3D features” for COCO images
- Final input features were always concatenation of individual features
- Benefits of “fake I3D features” for the model training:
    - We can use also 414113 COCO captions
    - We can use genuine I3D features of TGIF and VATEX

A selection of results on VTT 2019 ground truth data

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		0.2263	0.2812	0.1667	0.0446
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

TRECVID 2019 result

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		0.2263	0.2812	0.1667	0.0446
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

Adding VATEX dataset to COCO and TGIF improves

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		**0.2071**	**0.2746**	**0.1610**	**0.0443**
X	X	X	X	X	X		0.2263	0.2812	0.1667	0.0446
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

Adding ResNext features to ResNet and I3D improves

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		**0.2263**	**0.2812**	**0.1667**	**0.0446**
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

Using COCO data with TGIF and VATEX improves

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		**0.2263**	**0.2812**	**0.1667**	0.0446
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

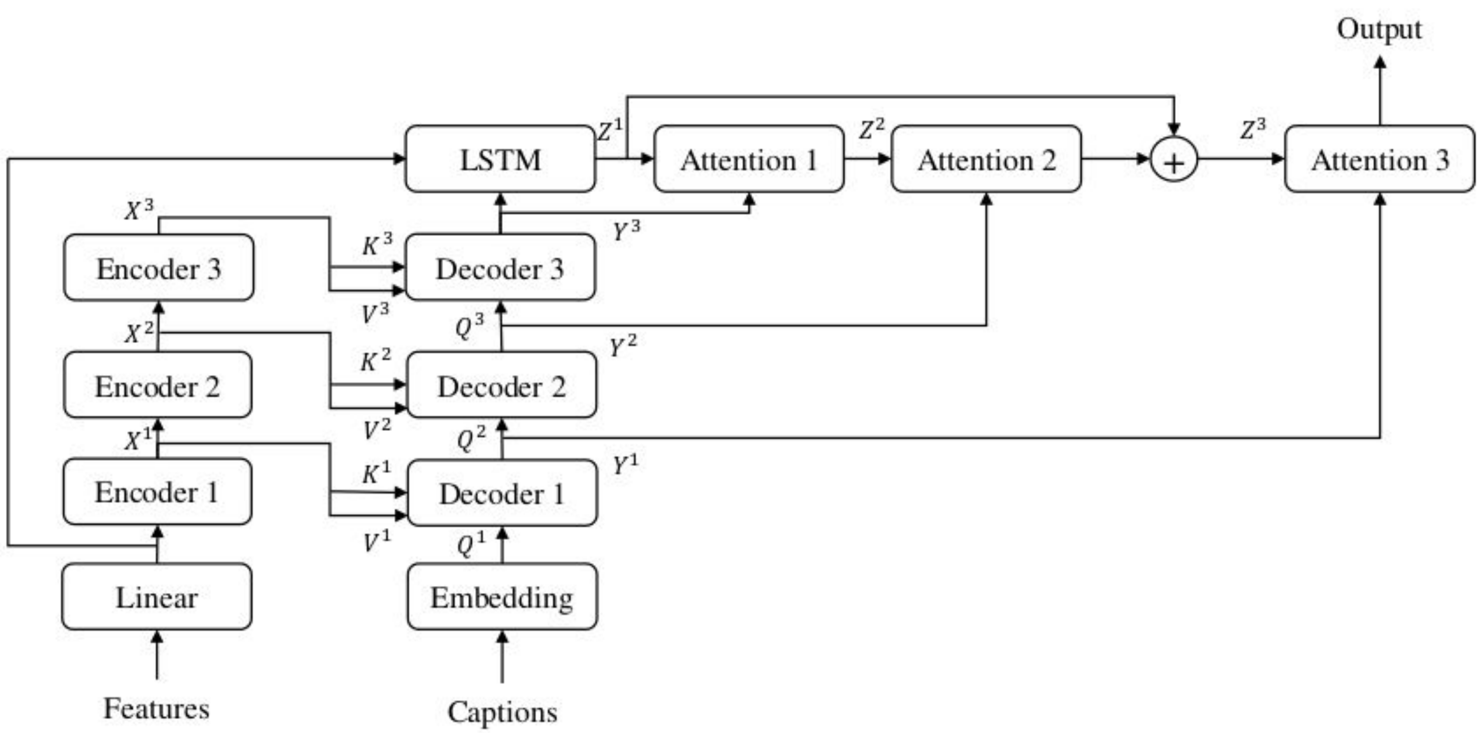
I3D is better than C3D as a video feature

CO CO	TG IF	VAT EX	Res Net	Res Next	I3D	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		0.2263	0.2812	0.1667	0.0446
	X	X	X	X	X		**0.2253**	**0.2528**	**0.1518**	**0.0446**
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

Final selection of datasets and features

COCO	**TGIF**	**VATEX**	**ResNet**	ResNext	**I3D**	C3D	METEOR	CIDEr	CIDErD	BLEU-4
X	X		X		X		0.2049	0.2348	0.1147	0.0319
X	X	X	X		X		0.2071	0.2746	0.1610	0.0443
X	X	X	X	X	X		0.2263	0.2812	0.1667	0.0446
	X	X	X	X	X		0.2253	0.2528	0.1518	0.0446
	X	X	X	X		X	0.2151	0.2345	0.1364	0.0397

Stacked attention captioning model



Stacked attention captioning model

- Based on the Transformer attention model

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_{model}}})V$$

- Uses multihead attention

$$Multihead(Q, K, V) = concat(h_1, \ldots, h_k)W^O$$
$$h_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$
$$i = 1, \ldots, k\ ,$$

- Intra-modality stacked attention for visual features in the encoder
- Inter-modality stacked attention from visual to textual features in the decoder
- Stacked attention from decoder outputs to caption generation

$$StackedAttention(Y^{N-j+1}, Z^j) = \alpha(Y, Z) \odot Z \qquad \alpha(Y, Z) = \sigma(W\,[Y, Z] + b)$$

Submissions

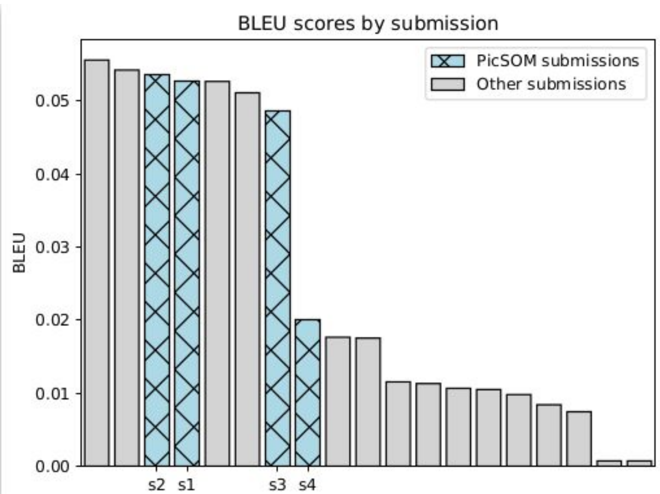
run	description
1	Our latest and best **stacked attention** model, trained with **COCO+TGIF+VATEX**
2	Model similar to our best **VTT 2019 submission**, trained with **COCO+TGIF+VATEX**
3	Another well-performing **stacked attention** model, trained with **COCO+TGIF**
4	Model similar to our best **VTT 2019 submission**, trained with **COCO+TGIF**

In all submissions cross-entropy training was used to initialize the LSTM model and self-critical reinforcement learning to finetune it with CIDEr-D score. TRECVID VTT 2018 ground truth data were used for validation.

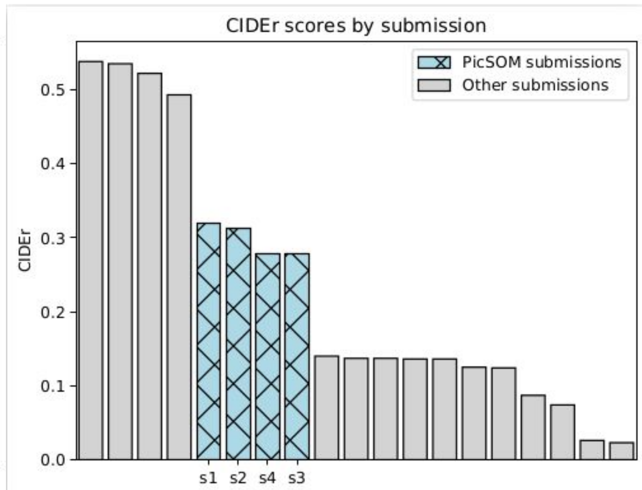
Results

run	METEOR	CIDEr	CIDErD	BLEU-4	SPICE	STS
1	**0.2617**	**0.319**	**0.200**	0.0527	**0.079**	0.4406
2	0.2556	0.312	0.191	**0.0536**	0.076	0.4293
3	0.2414	0.278	0.129	0.0485	0.069	**0.4581**
4	0.2323	0.278	0.124	0.0201	0.067	0.4458

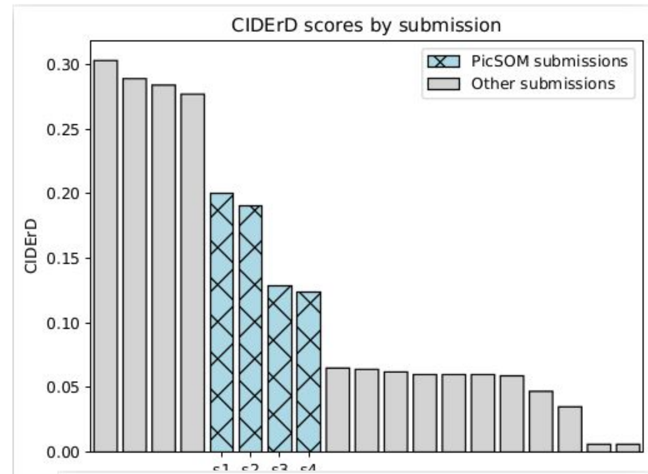
Plots



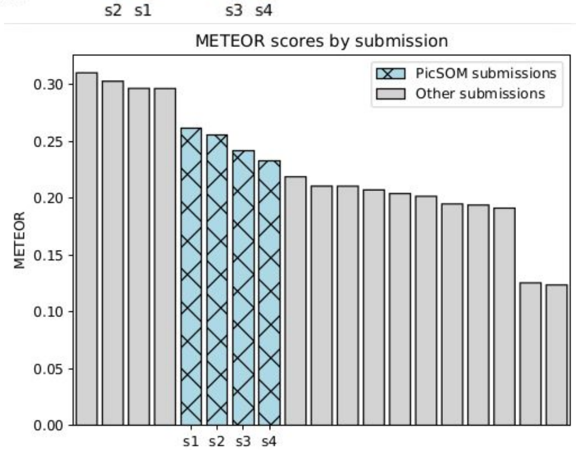
BLEU scores by submission
PicSOM submissions
Other submissions
BLEU
0.00
0.01
0.02
0.03
0.04
0.05
s2
s1
s3
s4



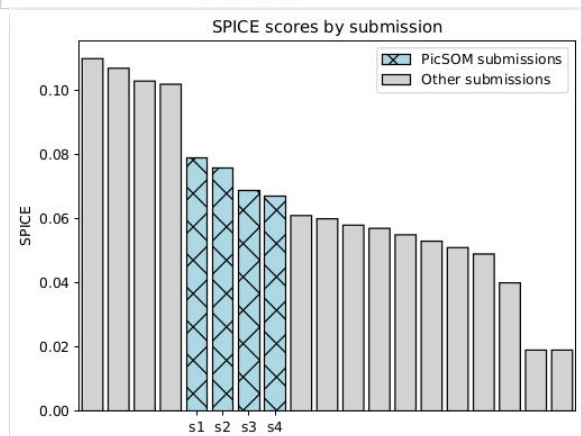
CIDEr scores by submission
PicSOM submissions
Other submissions
CIDEr
0.0
0.1
0.2
0.3
0.4
0.5
s1
s2
s4
s3



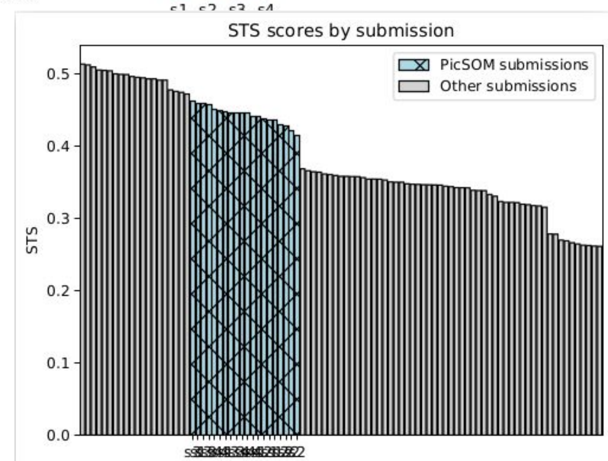
CIDErD scores by submission
PicSOM submissions
Other submissions
CIDErD
0.00
0.05
0.10
0.15
0.20
0.25
0.30



METEOR scores by submission
PicSOM submissions
Other submissions
METEOR
0.00
0.05
0.10
0.15
0.20
0.25
0.30
s1
s2
s3
s4



SPICE scores by submission
PicSOM submissions
Other submissions
SPICE
0.00
0.02
0.04
0.06
0.08
0.10
s1
s2
s3
s4



STS scores by submission
PicSOM submissions
Other submissions
STS
0.0
0.1
0.2
0.3
0.4
0.5

Observations & issues

- Using VATEX data was more beneficial than the stacked attention model
- Current version of stacked attention didn’t allow to use also ResNext features
- Our implementation of self-critical reinforcement learning is deficient as it creates captions in which the last word of the sentence is lacking
- Finding and correcting the bug will raise our scores substantially...