Aalto University
School of Science

Image Data, Video Data and Both in VTT Model Training

Video-to-Text Task in TRECVID 2019

Jorma Laaksonen, PicSOM Team

Department of Computer Science
Aalto University School of Science
Espoo, Finland

November 13th, 2019

Contents

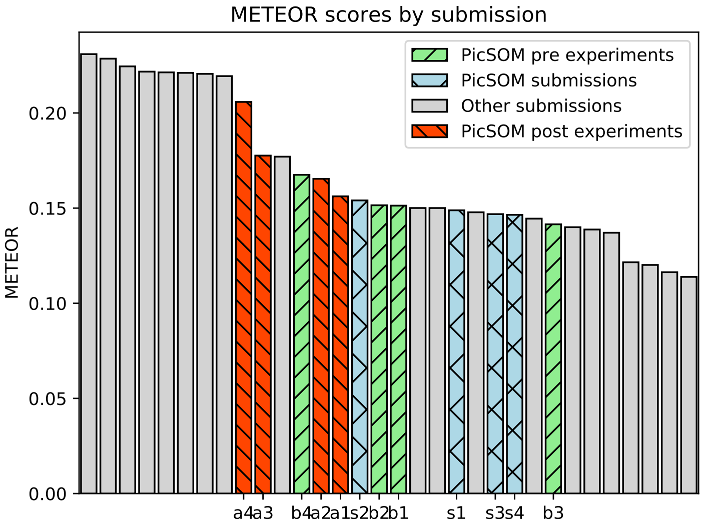
People

- Jorma Laaksonen
- Héctor Laria Mantecón
- (Danny Francis & Benoit Huet of EURECOM)

Lessons from TRECVID 2018

- We used only cross-entropy training, others did better with reinforcement learning
- Validation with VTT 2016 data was not able to select the best models
- Training with COCO image dataset gave equally good results as with video datasets
- We could move from old Theano-based code to new PyTorch-based

Development of scores



Work between TRECVID 2018 and 2019

- Implemented self-critical reinforcement learning
- Studied methods to combine image and video datasets and features
- Also wanted to study optimal combination of different video datasets

Contents

TGIF and COCO datasets

Statistics:

- TGIF: 125,713 videos with 125,713 captions
- COCO: 123,287 images with 616,767 captions

Which approach would be the best:

- 125,713 video feature vectors and 125,713 captions
- 123,287 image feature vectors and 616,767 captions
- 249,000 image feature vectors and 742,480 captions
- 249,000 image **and video** feature vectors and 742,480 captions

Videos to image features and vice versa

- Image features can be extracted from videos in multiple ways, e.g.
  - use only the middle frame
  - max or mean pool features of multiple or all frames
- Genuine video features such as I3D cannot be extracted from still images
  - we used **fake video features** for COCO images
  - **average of all video features in TGIF** was used assigned to all COCO images
- The final feature vector was concatenation of

TGIF videos:	I3D video feature	ResNet image feature of middle frame
COCO images:	constant average I3D feature	ResNet image feature

Contents

Methodology

- COCO image and TGIF video datasets in training
- model validation and early stopping with VTT 2018 dataset
- ResNet-152 CNN image and I3D video features
- fake I3D video features for COCO images
- "DeepCaption" LSTM language model decoder in PyTorch
- cross-entropy loss training in the beginning
- self-critical reinforcement learning in the end

Submissions

We submitted four runs:

- PicSOM.1-MeMAD.primary: uses ResNet and I3D features for initialising the LSTM generator, and is trained on MS COCO + TGIF using self-critical loss,
- PicSOM.2-MeMAD: uses I3D features as initialisation, and is trained on TGIF using self-critical loss,
- PicSOM.3: uses ResNet features as initialisation, and is trained on MS COCO + TGIF using self-critical loss,
- PicSOM.4: is the same as PicSOM.1-MeMAD.primary except that the loss function used is cross-entropy,

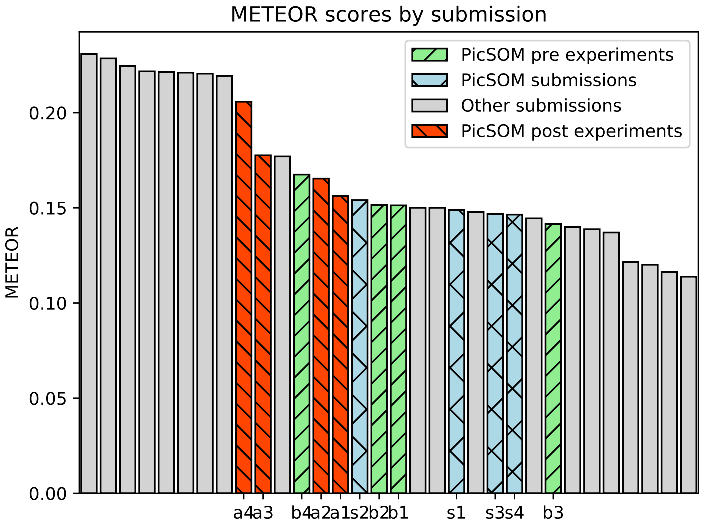
Contents

Results

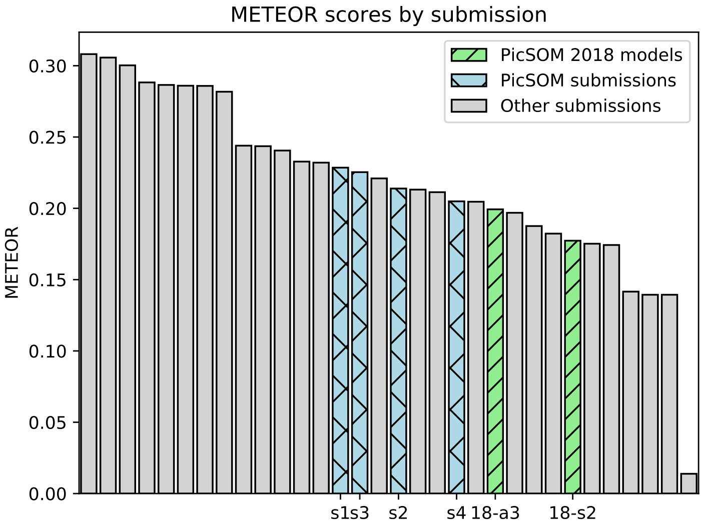
id	setup				2018				2019				
	t	**loss**	**feat**	**data**	**METEOR**	**CIDEr**	**CIDErD**	**BLEU**	**METEOR**	**CIDEr**	**CIDErD**	**BLEU**	**STS**
p-18-s2	I	ce	rn+fr	C+M	0.1541	0.1657	0.0476	0.0091	0.1773	0.1858	0.0722	0.0207	–
p-18-a3	I	ce	rn	C+T	0.1776	0.1948	0.0700	0.0197	0.1993	0.2174	0.1004	0.0288	–
p-19-s1	B	sc	rn+i3d	C+T	**0.2055**	**0.3025**	**0.1157**	0.0294	**0.2285**	**0.3277**	**0.1615**	**0.0385**	0.4168
p-19-s2	V	sc	i3d	T	0.1958	0.2718	0.0949	**0.0348**	0.2139	0.2773	0.1245	*0.0379*	*0.4169*
p-19-s3	I	sc	rn	C+T	*0.2007*	*0.2777*	*0.1074*	*0.0301*	*0.2254*	*0.3130*	*0.1569*	0.0345	**0.4282**
p-19-s4	B	ce	rn+i3d	C+T	0.1850	0.2190	0.0822	0.0213	0.2049	0.2348	0.1147	0.0319	0.4057

- p-18-s2 is our best submission in TRECVID 2018
- p-18-a3 is our best TRECVID 2018 post-conference result
- p-19-s* are our TRECVID 2019 submissisons

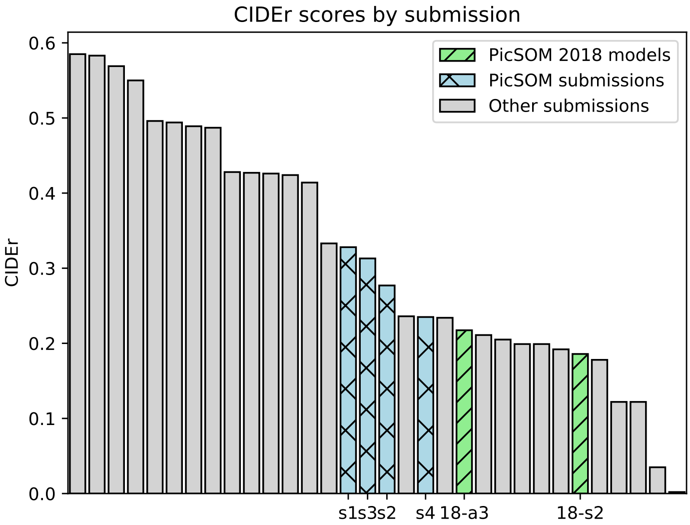
Comparison: METEOR 2018



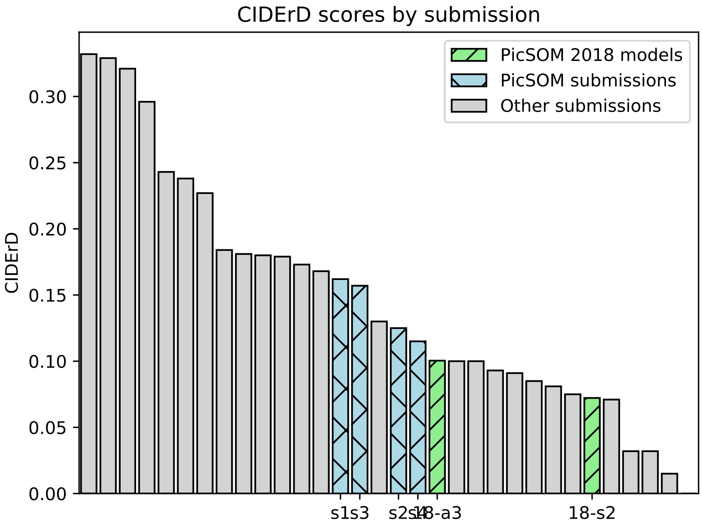
Comparison: METEOR



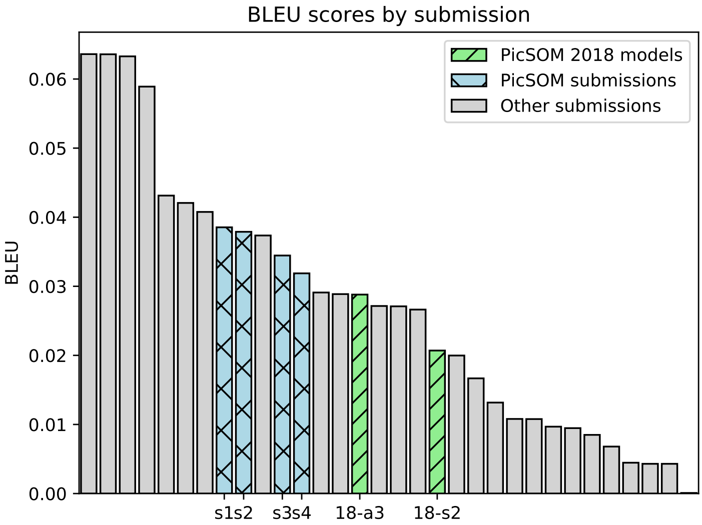
Comparison: CIDEr



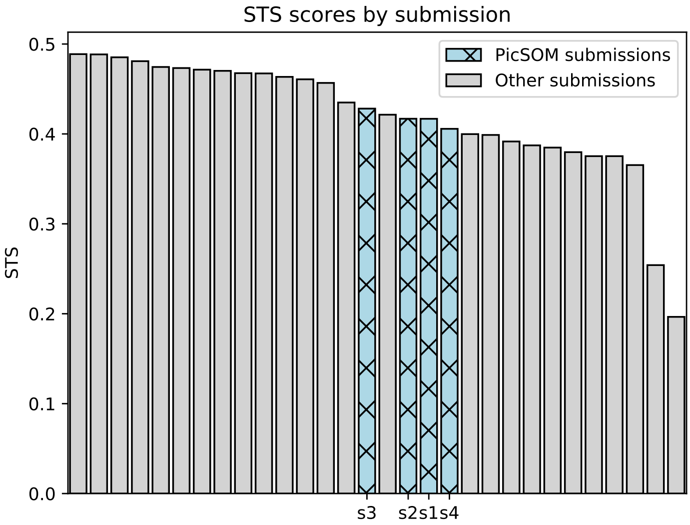
Comparison: CIDEr-D



Comparison: BLEU-4



Comparison: STS



Comparison

- s4 run is always the worst — reinforcement learning is beneficial
- s1 run is almost always the best — combining image and video features is good
- s3 run wins s2 with 4–1 — COCO image features better than TGIF video features

Contents

Run types

In TRECVID VTT 2019 all submissions had to be tagged with their run type:

- Run type ’I’: Only image captioning datasets were used for training
- Run type ’V’: Only video captioning datasets were used for training
- Run type ’B’: Both image and video captioning datasets were used for training

Run types per team

team	image	video	both
EURECOM		1	
FDU		2	
IMFD_IMPRESEE		3	
Insight_DCU			1
KU_ISPL		3	
KsLab		4	
PicSOM	1	1	2
RUCMM		4	
RUC_AIM3		4	
UTS_ISA		4	
10 teams	1	26	3

Training datasets used per team

team	COCO	TGIF	MSR-VTT	MSVD	VTT	VATEX	
EURECOM		X	X	X			0+3
FDU		X					0+1
IMFD_IMPRESEE			X				0+1
Insight_DCU		X					0+1
KsLab		X			X		0+2
PicSOM	X	X					1+1
RUCMM		X	X	X			0+3
RUC_AIM3		X	X		X	X	0+4
UTS_ISA		X	X	X	X		0+4
9 teams	1	8	5	3	3	1	0+0

Statistics of the training datasets

dataset	items	captions
COCO	123,287 img	616,767
TGIF	125,713 vid	125,713
MSR-VTT	6,513 vid	130,260
MSVD	1,969 vid	80,800
VTT	3,753 vid	9,020
VATEX	41,300 vid	826,000
LSMDC	108,536 vid	108,536

Video features used per team

team	I3D	C3D	CNN+pool	CNN+seq	audio
EURECOM	X				
FDU				X	
IMFD_IMPRESEE	X	X			
Insight_DCU		X			
KsLab			X		
PicSOM	X				
RUCMM		X	X		
RUC_AIM3	X			X	X
UTS_ISA	X			X	
9 teams	5	3	2	3	1

Contents

Conclusions

- In the PicSOM experiments the use of also the COCO dataset proved to be beneficial
- Naïve use of fake video features for images was better than not to use images at all
- This conclusion might be different if
  - our overall result level were higher
  - we used more video data than just TGIF
  - we used better video features than I3D
  - we used pooling or RNN based aggregation of framewise features
  - our implementation of self-critical training were better
- Model performance was very stable from validation with 2018 data to 2019 test data
- No other team used COCO dataset anymore
- Our results we clearly behind those of the best teams
- Specifying the run types in the way it was done now might be discontinued