INF@TRECVID2017 Video to Text Description

Jia Chen[1], Shizhe Chen[2], Qin Jin[2], Alexander Hauptmann[1]

Carnegie Mellon University[1]

Renmin University of China[2]

Main focus in this year: cross-dataset generalization

- Last year:
  - As the video caption pilot task provides no training captions for videos, we treat it as an *opportunity* to test the generalization ability of the caption models.
- This year:
  - We found that the performance of caption model begins to saturate within one dataset by comparison to human reference
  - opportunity->problem that we must face now

Motivation

- human reference on MSRVTT
  - leave-one-out test on groundtruth
- on par with the human reference on caption metrics
  - metric issue?
  - dataset issue (coupling with generalization issue)?

model	BLEU@4	METEOR	CIDEr
TGM	45.41	29.73	52.91
human	53.15	29.77	50.23

Motivation

- eliminate the metric issues
- on par with the human reference on tagging metrics (stop words removed)

model	precision	recall	f1
MP	77.4	17.2	26.8
tagging (top5)	47.8	12.5	18.6
tagging (top10)	38.6	17.1	22.2
human	70.7	20	29.7

Motivation

- preliminary cross dataset expriment

train	BLEU@4	METEOR	CIDEr
MSVD	47.70	34.22	80.88
MSR-VTT	34.67	30.68	55.39

- pitfall in the dataset MSRVTT:
  - train/test clips could come from the same video
  - The median number of shots for single video clip is 2 in MSRVTT
  - information leakage
- MSVD
  - too few videos
  - too many duplicate groundtruth sentences, which reduce the number of unique (video, caption) pairs

Cross-dataset Generalization Property of Models

- Q1: Which one is more promising for better generalization on unseen datasets, higher quality training dataset or more robust model?
- Q2: Could we get more stable generalization ability by ensembling more different models?

Basic Setting

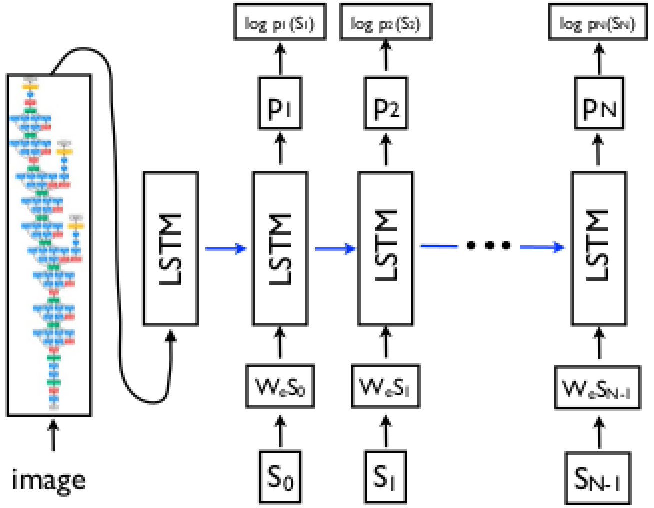
- Feature:
  - resnet200
  - i3d
  - mfcc (bow + fv)
- RNN with LSTM Cell
  - 512 hidden dimension, 512 input dimension
- Train scheme
  - batch size of 64

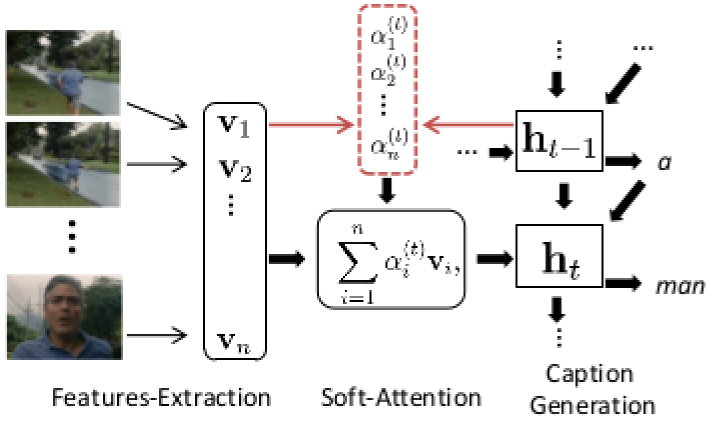
Q1: Higher quality training dataset or more robust model for better generalization?

- fix the model architecture to study its influence by treating TRECVID2016 as unseen dataset
- fix the training datasets to study its influence by treating TRECVID2016 as unseen dataset

Q1: Higher quality training dataset or more robust model for better generalization?

- Models:
  - Vanilla Encoder-decoder (MP)
  - Attention Encoder-decoder (ATT)
- Training dataset:
  - MSRVTT+MSVD
  - TGIF





Q1: Higher quality training dataset or more robust model for better generalization?

- the performance gain from dataset >> the gain from the caption model

Table 2: Comparison of changing model and change training sets

model	train dataset	BLEU4	Meteor	Cider
MP	MSRVTT+MSVD	5.04	12.13	30.25
ATT	MSRVTT+MSVD	5.59	12.38	31.96
MP	TGIF	8.05	14.67	37.00
ATT	TGIF	7.93	14.65	37.11

Q1: Higher quality training dataset or more robust model for better generalization?

- TGIF Dataset collection instruction:

DOs

- Please use only English words. No digits allowed (spell them out, e.g., three).
- Each sentence must contain between 8 and 25 words. Try to be concise.
- Each sentence must contain a verb.
- If possible, include adjectives that describe colors, size, emotions, or quantity.
- Please pay attention to grammar and spelling.
- Each sentence must express a complete idea, and make sense by itself.
- The sentence should describe the main characters, actions, setting, and relationship between the objects.

DONTs

- The sentence should **NOT** contain any digits.
- The sentence should **NOT** mention the name of a movie, film, and character.
- The sentence should **NOT** mention invisible objects and actions.
- The sentence should **NOT** make subjective judgments about the GIF.

Q2 Could we get more stable generalization ability by ensembling more different models?

- more replicas of models:
  - varying the detailed settings such as tuning dropout rate and using different epochs in training
- ensemble:
  - rerank sentences using the submitted model in the retrieval subtask

Q2 Could we get more stable generalization ability by ensembling more different models?

- by ensembling more and more models from source domain datasets, the performance on the target domain dataset TRECVID16 improves consistently

Table 3: Performance of ensembling

model	BLEU4	Meteor	Cider
best single model	8.05	14.67	37.00
ensemble 5 models	8.25	14.94	38.39
ensemble 6 models	8.25	15.04	38.66
ensemble 7 models	8.31	14.99	39.15
ensemble 8 models	8.46	15.04	40.79

Challenge Result

rank	mean.cider		bleu.ref2	
1	**RUC_CMU.run1.primary**	0.437	**RUC_CMU.run3**	0.022698561
2	**RUC_CMU.run2**	0.414	**RUC_CMU.run1.primary**	0.022503469
3	**RUC_CMU.run3**	0.411	**RUC_CMU.run2**	0.021839473
4	mediamill_generation_rerank	0.355	VTT17_Generation_Task_Team_INF_vtt16tuned.primary	0.015388222
5	**RUC_CMU.run4**	0.331	**RUC_CMU.run4.txt**	0.014392141

rank	meteor.ref2		sts.ref1	
1	**RUC_CMU.run1.primary**	0.198482183	**RUC_CMU.run1.primary**	0.461612502
2	**RUC_CMU.run2**	0.195623761	**RUC_CMU.run2.txt.1.sts**	0.455437854
3	**RUC_CMU.run3**	0.195056582	mediamill_generation_baseline	0.452634668
4	mediamill_generation_resnext_rerank_places2	0.178886646	**RUC_CMU.run3.txt.1.sts**	0.452282212
5	mediamill_generation_priority_run.primary	0.178122645	mediamill_generation_rerank	0.450247801