Mining for meaning: from vision to language through multiple networks consensus

Iulia Duță[1,2], Andrei Nicolicioiu[1,3], Vlad Bogolin[4], Marius Leordeanu[1,3,4]

iduta@bitdefender.com, anicolicioiu@bitdefender.com, vladbogolin@gmail.com, marius.leordeanu@imar.ro

[1]Bitdefender, Romania [2]University of Bucharest, Romania
[3]University Politehnica of Bucharest, Romania [4]Institute of Mathematics of the Romanian Academy

Accepted at The British Machine Vision Conference (BMVC), 2018: http://bit.ly/mining-for-meaning

1. Introduction

We tackle the challenging task of describing the content of video in natural language, called video captioning.

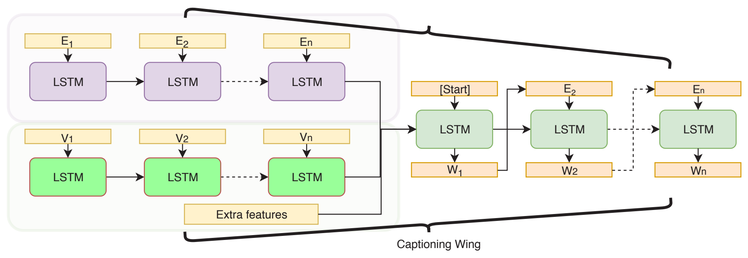
Our approach:

- generate a caption by means of consensus between the results of different models
- obtain diversity in the generated sentences by:
  - varying the video encoder
  - use sparse intermediate representations
  - leverage learning on additional tasks
- train 4 different network architectures: seq2seq, two-wings, two-stage, tcn
- use a selection method based on agreement and on pairwise comparisons between sentences

Main Contributions:

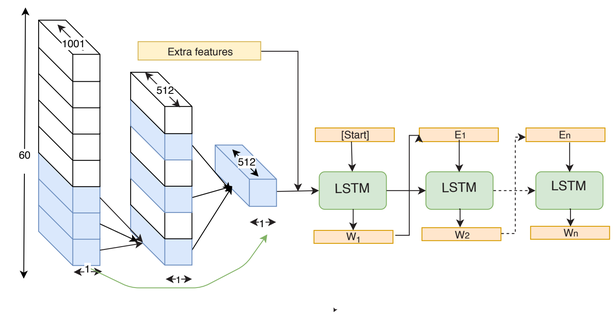
- propose **a method for selecting** a sentence that best describes a video
- propose **two novel architectures** and perform extensive tests with many others adapted from the literature
- achieve **state of the art** results on the MSR-VTT dataset.

2. Two-Wings Network



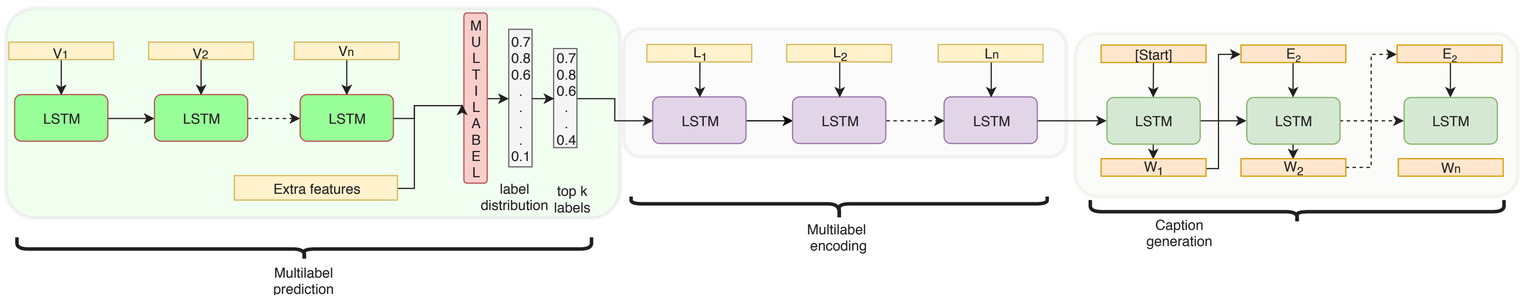
- **goal:** improve vocabulary of generated sentences
- improve language decoder by optimizing also for an auxiliary task
- learn a separate branch for **language reconstruction** on raw text - Wikipedia

3. TCN



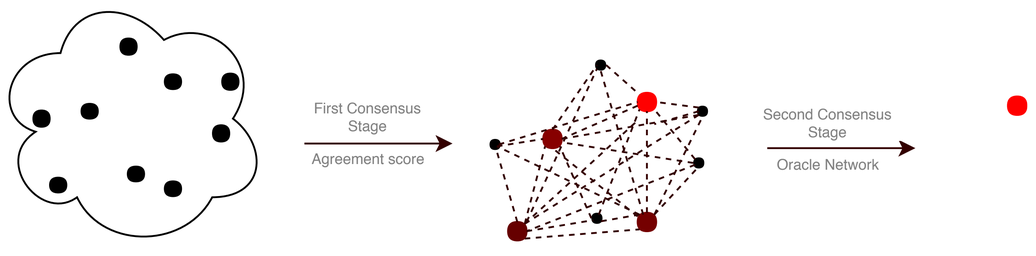
- **goal:** obtain a different video encoding
- use **temporal convolution** to aggregate features from neighbouring time steps
- encode the information through a hierarchy of convolutional layers into a single vector

4. Two-Stage Network



- **goal:** use sparse representation of the video
- fist stage: learn to **predict labels** from video
- second stage: learn to construct sentences from a set of labels
- learn models separately then fine-tune them jointly

5. Consensus



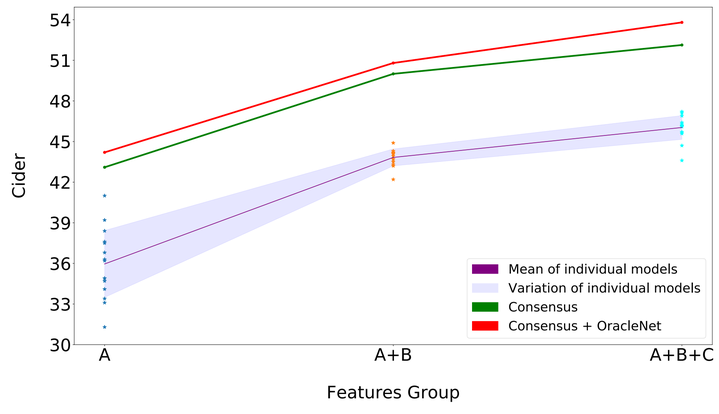
First consensus score:

- select the sentences that agree most with the others
- agreement score: for each generated sentence, compute its **CIDEr score against the others**
- choose the top C sentences

Oracle Network:

- train **a network to choose between 2** sentences given a video
- pairwise comparisons between top C and all the rest and count the number of wins for every sentence in top C
- final caption is the one with most wins

6. Features



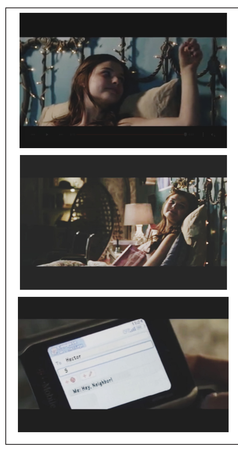
- each additional set of features bring improvement compared to single model
- consensus brings substantial improvements regardless of features used

4. Results

	CIDEr	Meteor	Rouge	Bleu 4
v2t navig [1]	44.8	28.2	60.9	40.8
MT-Ent [2]	47.1	28.8	60.2	40.8
HRL [3]	48.0	28.7	61.7	41.3
dense [4]	48.9	28.3	61.1	41.4
CIDEnt-RL [5]	51.7	28.4	61.4	40.5
TGM [6]	52.9	**29.7**	-	**45.4**
Ours	**53.8**	**29.7**	**63.0**	44.2

We obtain **state of the art** results on three evaluation metrics on MSR-VTT 2016 test set.

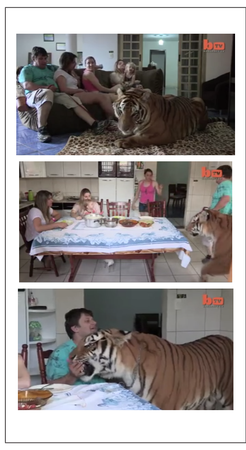
6. Qualitative Results



Top captions sorted by agreement score
4.660. a girl is knocking on the wall and texting
4.103. a girl is knocking on the wall
3.601. a girl is knocking on a wall and texting
3.359. a girl is knocking on her phone
3.119. a girl is knocking on her bed

Top captions sorted w.r.t. ground truth
1.518. a girl is knocking on the wall
1.315. a girl is knocking on the wall and texting
1.310. a girl laying in bed and knocking on the wal
1.289. a girl is laying in bed and knocking on the wall
1.109. a girl is knocking on a wall and texting

Human annotations
a girl in bed
a girl is knocking on the wall
a girl knocking on a wall
a girl knocks on a wall and texts a friend
a girl lays in bed and uses her phone



Top captions sorted by agreement score
2.321. a man and a woman are sitting in a table
1.118. a man is eating food from a table
1.070. a man is sitting in a table with a big dog
1.069. a group of people are sitting in a line with a tiger
0.925. a man is sitting in a chair with a tiger

Top captions sorted w.r.t. ground truth
0.650. a group of people are sitting in a line with a tiger
0.599. a man is sitting in a chair with a tiger
0.588. a man and a woman are eating a tiger in a bowl
0.547. a man is talking about a tiger
0.180. a man and a woman are sitting in a table

Human annotations
a story about a family that has seven tigers
a family rearing tigers and feeding them in the home
a family and their children are sitting
at a table playing with a tiger
five people sitting on a couch and a tiger
laying by their feet'
a family in brazil has 7 tigers that live in the house as petst

7. References

[1] Jin et al., ACM MM 2016 [2] Pasunuru and Bansal, ACL 2017 [3] Wang et al., CVPR 2018 [4] Shen et al., CVPR 2017 [5] Pasunuru and Bansal, EMNLP 2017 [6] Jin et al., ACM MM 2017