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

http://bit.ly/mining-for-meaning

1. Overview

Video captioning: describe videos in natural language

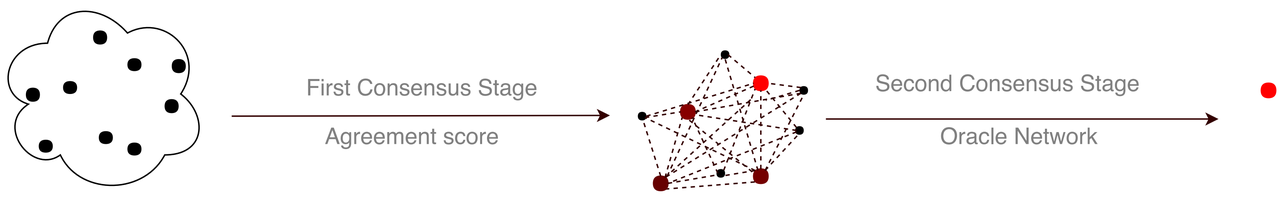
Our approach:

- obtain a diverse pool of generated sentences by:
  - varying the **video encoder** (TCN)
  - use **sparse intermediate representations** (Two-Stage)
  - leverage learning on **additional tasks** (Two-Stage, Two-Wings)
- use a selection method based on:
  - consensus among **whole pool** of sentences for a video
  - **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. Consensus



Agreement score:

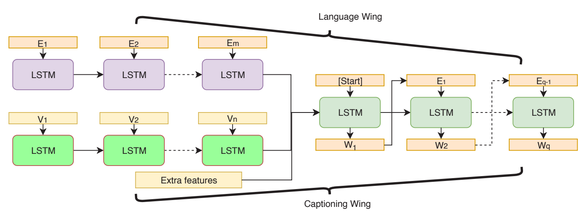
- select the sentences that agree most with the others
- agreement score: for each generated sentence, compute its **CIDEr score against the others**
- choose the top C sentences

Oracle Network:

- train **a network to choose between 2** sentences given a video
- pairwise comparisons between each sentence from top C and all the others from the pool
- final caption is the one with most wins

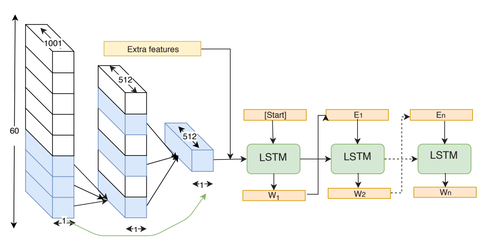
3. Architectures

Two-Wings



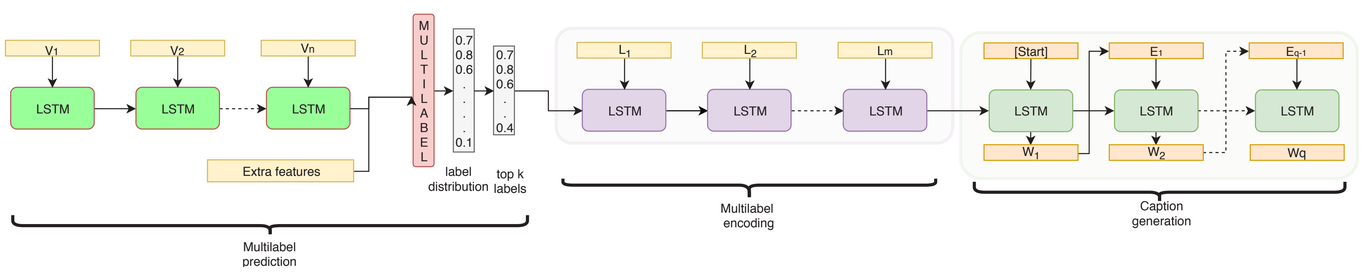
- *goal:* improve vocabulary of generated sentences
- improve language decoder by also learning a **language reconstruction task**
- use a separate branch (**shared decoder**) for optimizing on raw text - Wikipedia

TCN



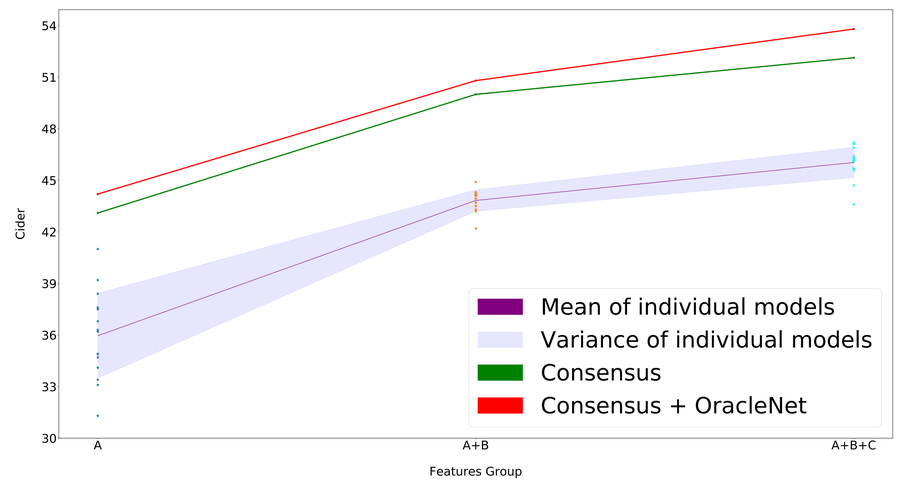
- *goal:* obtain a different video encoding
- use **temporal convolution** to aggregate features from neighbouring time steps
- encode the information - hierarchy of dilated convolutional layers

Two-Stage Network



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

4. Features



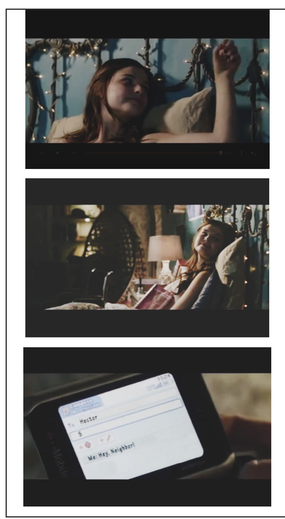
- each **additional set of features** bring **improvement** compared to single model
- **consensus** brings substantial **improvements** regardless of features used

5. 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

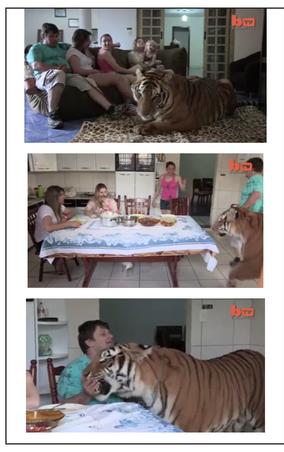
We obtain **state of the art** results on three evaluation metrics on MSR-VTT 2016 test set.

6. Qualitative Results



Top generated captions:
a girl is knocking on the wall and texting
a girl laying in bed and knocking on the wal
a girl is laying in bed and knocking on the wall
a girl is knocking on a wall and texting

Human annotations:
a girl in bed
a girl knocking on a wall
a girl lays in bed and uses her phone



Top generated captions:
a group of people are sitting in a line with a tiger
a man is sitting in a chair with a tiger
a man is talking about a tiger
a man and a woman are sitting in a table

Human annotations:
a story about a family that has seven tigers
five people sitting on a couch and a tiger laying by their feet

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