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

Iulia Duţă*[12]
iduta@bitdefender.com
Andrei Liviu Nicolicioiu*[13]
anicolicioiu@bitdefender.com
Simion-Vlad Bogolin*[4]
vladbogolin@gmail.com
Marius Leordeanu[34]
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

1 Qualitative results

In this Section we present some additional qualitative results. Firstly, we give some results produced by our pool of models and show the top sentences when they are ranked by the score with respect to the ground truth versus when they are ranked by the consensus score. Ideally, the consensus score should produce a ranking as close as possible to the ranking produced by the comparison to the ground truth. We then present results from the language reconstruction part of the Two-Wings Network and show that it is effective in producing coherent sentences from unordered sets of words. In the final part we display the intermediate, predicted word labels by the Two-Stage Network to better understand their relation to the final caption produced.

Consensus vs. Ground Truth Ranking: In Tables 1, 2, 3 and 4 we show some qualitative examples of sentences generated by our 16 models from multiple videos. On the left column of the tables we present 4 frames sampled from each video. The right side is split in 3 cells: in the upper cell we present top 5 generated sentences sorted by consensus score (with actual value shown at the start of each line), in the middle cell we list the top 5 sentences ordered by CIDER score with respect to ground truth (actual value shown at the start of each line) and in the bottom cell we randomly sample 5 examples of from human annotations.

Notice that our models produce meaningful and coherent sentences. In general we observed that the rank of a sentence in the order given by the consensus score is strongly correlated with the true rank in the order given by the ground truth score. This fact can also be seen in the examples presented below. Thus, the top scoring consensus sentences are also top scores with respect to ground truth. Therefore the consensus is a powerful automatic ranking scheme that could be reliably used to select good quality, meaningful sentences. The top-C

*The first three authors contributed equally.

captions are likely to contain top sentences with respect to ground truth, from which the Oracle Network can make the final selection. Also note that the human annotations, while generally agreeing on content, have varied degrees of fluency and quality.

Table 1: Example results generated by the pool of our 16 models. We present sentences ordered by consensus (first row) and by CIDEr score computed w.r.t. ground truth (second row) and human annotation (final row).

	**Top generated sentences ordered by consensus scores** 2.798. two men are playing tennis on a court 2.451. two men are playing tennis on a tennis court 2.065. a tennis player in blue and blue shirt is playing tennis 2.041. two men are playing a tennis game 1.938. a tennis player is hitting a ball in a match **Top generated sentences ordered by CIDEr score w.r.t. ground truth** 1.670. two men are playing tennis on a court 1.577. a tennis match is being played between two players 1.443. two men are playing a tennis game 1.356. two men are playing tennis on a tennis court 1.296. a tennis match between two men in blue and blue shirt **Human annotations** a tennis match is being played between two men two men participate and play in a tennis match two people playing in a tennis match a tennis match between two men with an advertisement for rolex a tennis piont ends with one player s signature shot
	**Top generated sentences ordered by consensus score** 3.287. a group of people are watching a fireworks display 3.179. a large crowd of people are watching a fireworks display 2.731. a crowd of people are watching a fireworks show 2.689. a large fireworks display is being shown 2.440. a large fireworks display **Top generated sentences ordered by CIDEr score w.r.t. ground truth** 1.536. a group of people are watching a fireworks display 1.527. a crowd of people are watching a fireworks show 1.330. a group of fireworks are going to the crowd 1.325. a large crowd of people are watching a fireworks display 1.209. a fireworks display is going off **Human annotations** a clip showcasing fireworks going off in the sky a crowd is cheering at fireworks a crowd of people are recording a fireworks show with their cellphones a crows takes photos of fireworks shooting from a building a group of people are watching fireworks

Table 2: Additional qualitative results of sentences ranked by consensus versus ground truth ranking.

Top generated sentences ordered by consensus 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 generated sentences ordered by CIDEr score 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

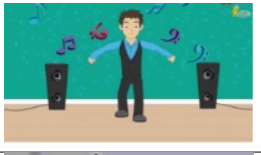
Top generated sentences ordered by consensus score
5.699. a person is opening a toy
4.238. a person is opening a toy with a toy
4.039. a person is opening a package
4.015. a person is opening a box
3.666. a person is opening a red package

Top generated sentences ordered by CIDEr score w.r.t. ground truth
2.131. a person is opening a toy
1.586. a person is opening a box
1.445. a person is opening a toy with a toy
1.441. a person is opening a package
1.367. a person is opening a red package

Human annotations
a clip of someone taking toys out of a gift set
a man is opening a toy egg
a man is playing with some toys
a man is showing off the cars gift set
an unboxing of some toys

Table 3: Additional qualitative results of sentences ranked by consensus versus ground truth ranking.

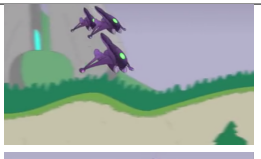


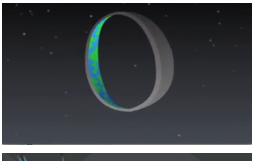



Top generated sentences ordered by consensus score
3.271. a cartoon character is singing
2.997. a cartoon character is singing a song
2.727. a cartoon of a man singing a song
2.705. a cartoon of a man singing and dancing
2.682. a cartoon of a girl singing and dancing

Top generated sentences ordered by CIDEr score w.r.t. ground truth
0.646. a cartoon of a man singing and dancing
0.627. a cartoon of a girl singing and dancing
0.479. a cartoon is singing
0.450. a cartoon character is dancing
0.423. a cartoon of a man singing a song

Human annotations
a kid s animated song
a animation band is singing song
a cartoon about the hokey pokey song and dance
a cartoon depicts the hokey pokey
a cartoon for children






Top generated sentences ordered by consensus score
3.380. a video game character is flying
3.170. a video game character is flying around
2.493. a cartoon character is flying
2.388. a cartoon character is flying a ball
2.268. a person is playing a video game

Top generated sentences ordered by CIDEr score w.r.t. ground truth
0.314. a cartoon of a man is flying around
0.269. a cartoon character is flying
0.243. a man is flying through a video game
0.233. a cartoon character is flying a ball
0.202. a cartoon character flying a monster

Human annotations
video game characters are flying through space
there were three characters flying in the air
cartoon characters sing about space
a video of halo the video game
halo cartoon animation music video

Table 4: Additional qualitative results of sentences ranked by consensus versus ground truth ranking.

Top generated sentences ordered by consensus 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 generated sentences ordered by CIDEr score 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 wall
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 generated sentences ordered by consensus score
1.991. a man is doing construction
1.907. a man is doing construction work
1.881. a man is doing construction improvement
1.855. a man is working on a floor
1.842. a man is installing a wood floor

Top generated sentences ordered by CIDEr score w.r.t. ground truth
1.132. a man is installing a wood floor
1.052. a man is installing flooring
0.867. a man is working on a wood floor
0.816. a man is fixing a wood floor
0.621. a man is doing flooring

Human annotations
a man is installing new flooring
a carpenter places down some wood floring
a man is decking a floor
a man is fixing the floor
a man is flooring

Language reconstruction results: In Table 5 we show results of the reconstruction part of the Two-Wings model. In this submodel we receive as input a sentence from the annotations, apply a random permutation on the order of the words, remove half of them and try to reconstruct it. The first column contains the sentence to be reconstructed, the second column the remaining shuffled words used as input and the third one the generated sentence.

We can see that the generated captions are grammatically and semantically correct. Although the reconstruction does not match the target sentence, given just half of the individual words and the random permutation of these words it would be impossible, even for a human, to reconstruct back the original sentence. But this is not the end goal of this branch - its main purpose is to learn to generate rich, diverse and coherent sentences. Consequently, in our experiments the Two-Wings Network produced on average more diverse sentences than other network models.

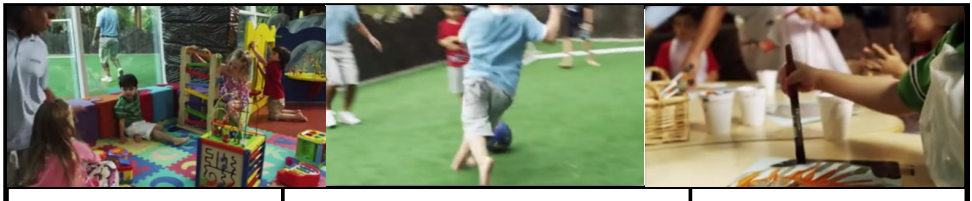
Table 5: Results of the language reconstruction part of the Two-Wings model. The first column contains the ground truth sentence, the second the broken input sentence and the last the generated sentence. We use this model to improve our language generating decoder.

Ground truth Sentence	Input words	Generated Sentence
a man talks about a young girl performing great in track	a about great performing young man a talks girl in	a man talking about a young girl in front of a crowd
someone giving demo in computer about something	in about computer someone demo giving	someone is showing video graphics on computer
a boy in a long sleeve shirt is playing the guitar	a boy the in long is sleeve a shirt playing	a young boy is playing in the water with a dog
a man is performing on stage with some other actors	stage with some on performing is man a	a man is singing with his band on stage
two guy: one in black tshirt is driving a car and other in white shirt is sitting next to him	is in and in sitting other car guy white tshirt two is a shirt driving black one	a man in a blue shirt is sitting and speaking with two other men while standing on a couch
a balding man in a suit gives a speech by flags	speech balding a suit in a a man gives	a man in a suit with a tie talking in a room
a baseball player in a red uniform while music plays	player in music red uniform a a baseball while	a man in a blue shirt and black shorts playing basketball

Two-Stage Network results: In Figure 1 we present some examples produced by the Two-Stage Model. The model consists of two parts: a part (a first stage) that predicts multiple labels for the video followed by a part (the next stage) that generates a sentence based only on these labels. The model is initialized by training both parts independently and then fine-tune them jointly.

For each cell, the first row contains 3 frames sampled from each video, the second row contains results from the initial independent learning and the third row contains final results after the fine-tuning. For each row, the first column shows a ground truth sentence, the second column shows the top K predicted labels with their corresponding probabilities and the last column shows the generated sentence.

animator explains the animation process	be(0.99) paper(0.99) man(0.99) person(0.99) sketch(0.99) someone(0.99) draw(0.99) show(0.99) guy(0.98) explain(0.98)	a person is showing a man who is trying to draw a wooden
a man is drawing characters using a pencil	be(0.99) draw(0.99) cursive(0.99) man(0.59) child(0.56) scissors(0.54) artist(0.53) chalkboard(0.52) tool(0.38) animate(0.32)	a man is drawing a cartoon character



a teacher is inside with children and kids are playing with soccer balls and paint	be(0.99) child(0.99) kid(0.99) video(0.99) play(0.98) people(0.98) man(0.98) show(0.97) (0.97) clip(0.97)	a play a video game is shown while pictures of young children are displayed
young kids play in a daycare	be(0.99) child(0.99) figurine(0.34) school(0.28) easter(0.28) pool(0.17) sidewalk(0.16) girl(0.13) chalkboard(0.10) plane(0.08)	a group of kids are playing with toys

women are protesting the lack of midwives	be(0.99) group(0.99) people(0.99) street(0.99) video(0.99) road(0.99) man(0.99) hold(0.99) walk(0.99) talk(0.98)	walking dead people are standing on a street holding a country
a group of people are protesting outside	be(0.99) sidewalk(0.99) child(0.87) ceremony(0.81) protester(0.77) soldier(0.70) girl(0.45) joking(0.28) heel(0.19) horse(0.12)	a group of people are walking down the street

Figure 1: Example of qualitative results of the Two-Stage Network. The first column contains a ground truth sentence, the second column the top 10 predicted labels and the third column contains the final generated sentence by the Two-Stage model. In top row of every box, we present the results of the model when we train the two parts separately. In the second row we show results of the two parts fine-tuned jointly, end-to-end on the MSR-VTT dataset.

Labels generated by our multi-label model have a high degree of accuracy. To improve the quality of the captions we fine-tuned the whole model end-to-end, obtaining significantly better results. While the multi-label prediction was better before fine-tuning, after the fine-tuning, which did not put a cost on the labels, the multi-label prediction decreased in accuracy, while the final quality at the caption level improved. This fact is also observed in the qualitative examples in Figure 1 as the fine-tuned model, trained end-to-end (third row) produces captions of better quality than the model with the two stages trained independently (second row). However, the fine tuned model is worse at predicting intermediate word labels - due to the end-to-end training with loss on the final caption but no intermediate loss on the intermediate multi-label prediction.