# End-to-End Video Captioning Supplementary material

Anonymous ICCV submission

Paper ID 3982

## 1. Supplementary Material

We show the additional materials about our work. In particular, we are going to add more qualitative results of our models, focusing on the difference between the two steps. Specifically, we analyse: (i) when step 2 improves over step 1; (ii) when step 2 does not improve the step 1; (iii) negative samples, i.e. when both the steps produce wrong captions.

The improvement of the model from step 1 to step 2 that we have found are summarised in the Tables 3, 4 and 5. As we expected, fine-tuning the model allows correcting the wrong words in the caption without changes the entire sentence (rows 1, 5 Table 3; rows 1, 2 Table 5). Also, it can add new information about the action, e.g. new words in the sentence (rows 5, 7, 8 in Table 5). Although, the second step improves most of the example of the test set, but, some of them show the opposite results. On the MSVD dataset (Table 1) it recurs in few elements, while in the MSR-VTT the model wrong to predict the subject or object of the action (rows 5, 7 in Table 2). Some negative example are showed in table 6 and 7. Many rows like 1, 2, 5, 6 in Table 6 and 1, 2, 3, 4 in Table 7 shows how to try the improvements of the model (step 2) in such cases does not work, in fact, the prediction does not change. On the other case (rows 10 in Table 6 and rows 5, 8 in Table 7) the changes just lead to mistakes in a different way. Finally, MSVD dataset allows to create a model that predicts a simple sentence, while the model trained using the MSR-VTT one creates longer sentences. Our framework is, therefore, able to provide reasonable predictions in both cases.

| EtENet-GLN | |
|---|---|
| Step 1 | **a man is eating something** |
| Step 2 | a man is eating a flute |
| References | *a man eats spaghetti*; *a man is eating something*; *a person is eating*; |
| Step 1 | **two women are dancing** |
| Step 2 | two men are fighting |
| References | *a group of people are dancing*; *several people are dancing on the patio*; *some people are dancing*; |
| Step 1 | **a baby is eating** |
| Step 2 | a baby is talking |
| References | *a baby is eating ice cream*; *a child is laughing*; *a baby holding ice cream is laughing*; |
| Step 1 | **a man is playing a piano** |
| Step 2 | a man is playing a guitar |
| References | *a man is playing a keyboard piano*; *a man is playing an electric keyboard*; *a man plays on the electronic musical instrument*; |
| EtENet-IRv2 | |
| Step 1 | **men are playing basketball** |
| Step 2 | men are playing soccer |
| References | *a basketball game is in play*; *two teams are playing basketball on television*; *men are playing basketball*; |
| Step 1 | **two men are fighting** |
| Step 2 | a man is talking |
| References | *two men are fighting*; *two men are fighting with each other*; *two men appear to be fighting or wrestling*; |
| Step 1 | **a man is playing cricket** |
| Step 2 | a man is playing football |
| References | *two teams are playing cricket*; *people are playing cricket*; *a man in cricket performs an error*; |

Table 1. Some qualitative examples in which the step 2 does not improve over the step 1 on the MSVD dataset. Bold text means the best results.

| EtENet-IRv2 | |
|---|---|
| Step 1 | **a man is playing a baseball game** |
| Step 2 | a man is running on a field |
| References | *a baseball game is played*; *players hitting baseballs with bat*; *a man is hitting the ball in a baseball game*; |
| Step 1 | **a group of people are dancing** |
| Step 2 | a group of people are walking down the street |
| References | *a dance class where people are learning footwork*; *a large class of men and woman taking dance lessons in a studio*; *a ballroom dance class*; |
| Step 1 | **a person is mixing ingredients in a bowl** |
| Step 2 | a woman is mixing ingredients in a bowl |
| References | *a woman mixes batter in a bowl*; *a chef stirs flower with water*; *a woman is saying how to make nabeyaki udon noodle*; |
| Step 1 | **there is a fish swimming in the water** |
| Step 2 | a fish swimming in a swimming pool |
| References | *two orange and white fish are swimming together*; *there are two fish floating in to the water*; *a fish tank with two gold fish and plants*; |
| Step 1 | **a baby is walking** |
| Step 2 | a turtle is walking |
| References | *a turtle is walking underwater*; *a turtle is swimming in water*; *the turtle is moving under water*; |
| Step 1 | **a man is talking to a crowd of people** |
| Step 2 | a group of people are walking on a stage |
| References | *a man calls out a young girl s name to walk onto the stage*; *a man is singing to a crowd*; *annual day celebrations are going on*; |
| Step 1 | **a group of people are playing basketball** |
| Step 2 | a man is running on a trampoline |
| References | *a group of young people play basketball together outside*; *people playing basketball and also performing trick shots and bloopers*; *a basketball player swats a ball when it is shot*; |
| Step 1 | **a man is talking about a car** |
| Step 2 | a man is explaining something |
| References | *a demonstration of a broken part to a vehicle*; *a guy talking about car parts*; *a man is repairing a car*; |

Table 2. Some qualitative examples with EtENet-IRv2 model in which the step 2 does not improve the step 1 using the MSR-VTT dataset. Bold text means the best results.

| EtENet-GLN | |
|---|---|
| Step 1 | a man is dancing |
| Step 2 | **a woman is dancing** |
| References | *a girl is dancing*; *a woman does aerobic exercise*; *a woman is exercising*; |
| Step 1 | a polar bear is floating on the water |
| Step 2 | **a polar bear is walking** |
| References | *a polar bear is running toward walruses*; *a polar bear is walking*; *bears are running*; |
| Step 1 | a man is cooking |
| Step 2 | **a man is pouring sauce in a pot** |
| References | *a man is pouring wine into a pot*; *a person is adding souse in the pot*; *a person making spaghetti sauce*; |
| Step 1 | a group of men are dancing |
| Step 2 | **a man is dancing** |
| References | *a man is dancing*; *a man is dancing on stage*; *a person is dancing*; |
| Step 1 | a woman is slicing a carrot |
| Step 2 | **a woman is slicing a tomato** |
| References | *a person is slicing a tomato*; *a woman is slicing tomato*; *a chef is slicing a tomato*; |
| Step 1 | a dog is walking |
| Step 2 | **a panda is playing** |
| References | *a baby panda is going down a slide*; *pandas are playing*; *panda babies are playing*; |
| Step 1 | a man is pouring sauce into a bowl |
| Step 2 | **a woman is adding water into a bowl** |
| References | *someone is pouring water into a plastic bowl of mushrooms*; *a woman pours some water on an unknown brown food*; *a woman is mixing ingredients*; |
| Step 1 | a man is dancing |
| Step 2 | **a man is riding a motorcycle** |
| References | *a man and woman ride a motorcycle*; *a lovers is riding on the motor bike*; *a man and woman are driving on a motorcycle*; |
| Step 1 | a band is singing |
| Step 2 | **a woman is singing** |
| References | *a woman is singing into a hand-held microphone on stage*; *a woman is singing*; *a girl is singing on stage*; |

Table 3. Some qualitative examples with EtENet-GLN model in which the step 2 improve the step 1 using the MSVD dataset. Bold text means the best results.

| EtENet-IRv2 | |
|---|---|
| Step 1 | a woman is slicing some bread |
| Step 2 | **a woman is slicing a potato** |
| References | *a woman is piercing potato*; *a lady is cooking food*; *a woman is stabbing a potato with a fork*; |
| Step 1 | a man is reading |
| Step 2 | **a man is talking** |
| References | *a man is talking*; *a man talks*; *a man is speaking directly to the camera*; |
| Step 1 | a person is mixing ingredients in a bowl |
| Step 2 | **a woman is mixing ingredients in a bowl** |
| References | *a woman mixes batter in a bowl*; *a chef stirs flower with water*; *a woman is saying how to make nabeyaki udon noodle*; |
| Step 1 | a woman is slicing a cucumber |
| Step 2 | **a woman is slicing an onion** |
| References | *a woman is slicing an onion*; *a woman cuts onion*; *a woman slices onions with a large knife*; |
| Step 1 | a baby is walking |
| Step 2 | **a turtle is walking** |
| References | *a turtle is walking underwater*; *a turtle is swimming in water*; *the turtle is moving under water*; |
| Step 1 | a woman is drawing a piece of paper |
| Step 2 | **a woman is cutting a piece of paper** |
| References | *a person is cutting a piece of paper*; *a woman is cutting papers*; *a woman is cutting a paper with a scissors*; |
| Step 1 | a man is playing with a man |
| Step 2 | **a man is playing a guitar** |
| References | *a man is playing the guitar*; *a man is sitting and playing a small guitar*; *the homeless guy played his ukelele on the street*; |
| Step 1 | a man and a woman are dancing |
| Step 2 | **a man is riding a boat** |
| References | *the woman is paddling a canoe*; *a woman is rowing a boat*; *a lady is rowing a boat*; |
| Step 1 | a person is cooking |
| Step 2 | **a person is cutting meat** |
| References | *a person is chopping meat*; *the person is slicing meat*; *a man is chopping beef using a large flat kitchen knife*; |

Table 4. Some qualitative examples with EtENet-IRv2 model in which the step 2 improve the step 1 using the MSVD dataset. Bold text means the best results.

| EtENet-IRv2 | |
|---|---|
| Step 1 | a person is playing a football game |
| Step 2 | **a man is talking about football** |
| References | *a man is talking about a sports match*; *guy speaking about matt ryan s new contract*; *a man talks about matt ryan while still frames are shown*; |
| Step 1 | a person is playing with toys |
| Step 2 | **a man is playing with a dog** |
| References | *a couple talk about their dog*; *a couple is shown with many pictures of their dog*; *people are talking and holding a dog*; |
| Step 1 | a person is mixing ingredients in a bowl |
| Step 2 | **a woman is mixing ingredients in a bowl** |
| References | *a woman mixes batter in a bowl*; *a chef stirs flower with water*; *a woman is saying how to make nabeyaki udon noodle*; |
| Step 1 | a man is talking |
| Step 2 | **a man is talking to another man** |
| References | *a man is talking to another man*; *a man talking to another guy*; *two men are talking in a dark room*; |
| Step 1 | a man is riding a boat in the ocean |
| Step 2 | **a man in a blue shirt is swimming in the ocean** |
| References | *there is a man in blue is swimming in the sea*; *there is a man in blue is talking nearby the beach*; *a man in blue shirt swims with big waves then gets on back of a jet ski*; |
| Step 1 | a woman is talking to a woman |
| Step 2 | **a woman is cooking food** |
| References | *man and woman on cooking show*; *a man stirs a bowl of mashed potatoes*; *a chef is making a dough*; |
| Step 1 | two men are playing table tennis |
| Step 2 | **a man in a blue shirt is talking about ping pong** |
| References | *a man describes a good ping pong stroke including keeping the paddle about head high*; *ping pong player explaining about some tricks to win matches*; *a left hand blue tshirt person is play table tennis*; |
| Step 1 | a man is singing |
| Step 2 | **a man and a woman are talking to each other** |
| References | *a man and a woman are having a conversation*; *woman is talking with man*; *woman talks about poetry and love*; |

Table 5. Some qualitative examples with EtENet-IRv2 model in which the step 2 improve the step 1 using the MSR-VTT dataset. Bold text means the best results.

| EtENet-GLN | |
|---|---|
| Step 1 | two girls are dancing |
| Step 2 | two girls are dancing |
| References | *a man dries off a woman*; *couples were speaking*; *a man is talking to a woman*; |
| Step 1 | a dog is walking |
| Step 2 | a cat is walking |
| References | *a guinea pig chews on food*; *the hamster is eating a carrot*; *a rabbit is eating a carrot*; |
| Step 1 | two men are playing |
| Step 2 | a man is riding a horse |
| References | *an elephant eats foliage*; *the elephant is eating*; *an elephant is eating grass*; |
| Step 1 | a band is performing on a stage |
| Step 2 | a man is playing with a stage |
| References | *the flag is waving in the air*; *the american flag flew in the wind*; *the united states flag is waving*; |
| EtENet-IRv2 | |
| Step 1 | a man is cutting a tomato |
| Step 2 | a man is cutting a tomato |
| References | *a man is putting a knife in a clamp*; *a man is balancing a knife*; *a man is keep the knife on the machine*; |
| Step 1 | a group of people are playing |
| Step 2 | a baby is eating |
| References | *a woman puts stickers on her face*; *the woman is putting stickers on their face*; *a girl is addicting stricers*; |
| Step 1 | two men are playing chess |
| Step 2 | a group of people are dancing |
| References | *a choir is singing*; *the peoples are singing a song*; *people are singing in church*; |
| Step 1 | a cheetah is running |
| Step 2 | a lion is running |
| References | *a jackal is walking around in a field*; *a jackal is running through grass*; *predator running in a jungle*; |
| Step 1 | a man is slicing a potato |
| Step 2 | a person is slicing a potato |
| References | *someone is using a juicer to squeeze the juice out of a lemon*; *a man is preparing sweet lemon juice*; *a man making orange juice*; |
| Step 1 | a group of men are playing |
| Step 2 | a man is doing some sort of bread |
| References | *someone opens a pizza box containing pepperoni pizza*; *a man opens a pizza box*; *a person is opening a pizza box*; |

Table 6. Some negative examples on MSVD dataset

| EtENet-IRv2 | |
|---|---|
| Step 1 | a man is holding a gun |
| Step 2 | a man is holding a gun |
| References | *someone repairing or assembling a machine*; *a man showing how to install grommets for a car engine*; *a man is teaching how to use tools*; |
| Step 1 | a man is playing a video game |
| Step 2 | a man is playing a video game |
| References | *a man is shooting a basketball ground*; *a photographer describes the space he used for a photo shoot*; *a photographer is taking photos*; |
| Step 1 | a woman is singing |
| Step 2 | a woman is dancing |
| References | *a woman trying to escape in a television scene*; *people are running on a sidewalk*; *people running and a woman falls down*; |
| Step 1 | a person is showing how to solve a toy |
| Step 2 | a person is showing how to solve a toy |
| References | *a person is working*; *a man is teaching how to tie a fishing knot*; *instructions for tying a knot*; |
| Step 1 | a man is playing a video game |
| Step 2 | a man is talking to a man |
| References | *filming of action scenes in movie*; *a man is filming a scene in the rain*; *a video showing the making of the movie thor*; |
| Step 1 | a cat is talking |
| Step 2 | a cat is talking to a cat |
| References | *a group of dogs are coming out of a container*; *man whistles and a dozen puppies come running out of a small round box*; *dogs are all sleeping together in a little room*; |
| Step 1 | a man is talking about a car |
| Step 2 | a man is talking about a car |
| References | *video game scene of a guy looking at different cars*; *cartoon play on the show*; *a man playing video games*; |
| Step 1 | a person is playing a video game |
| Step 2 | a man is talking about a car |
| References | *an animation presents something called project z*; *someone is showing video graphic*; *credits presented in 3d text*; |

Table 7. Some negative examples on MSR-VTT dataset