

**HW2 Video caption generation:**

## 1. Model summary:

```
In [65]: print(s2vt)
S2VT(
  (drop): Dropout(p=0.5, inplace=False)
  (linear1): Linear(in_features=4096, out_features=500, bias=True)
  (linear2): Linear(in_features=500, out_features=8748, bias=True)
  (lstm1): LSTM(500, 500, batch_first=True, dropout=0.5)
  (lstm2): LSTM(1000, 500, batch_first=True, dropout=0.5)
  (embedding): Embedding(8748, 500)
)
```

## 2. Data prepare:

trainset - List (1450 elements)

Index	Type	Size	Value
0	dict	2	{'caption': ['<BOS> A woman goes under a horse. <EOS>', '<BOS> A woman ...
1	dict	2	{'caption': ['<BOS> A man slicing butter into a bowl. <EOS>', '<BOS> A ...
2	dict	2	{'caption': ['<BOS> A raccoon-like animal is hanging upside down from t ...
3	dict	2	{'caption': ['<BOS> A man is putting pepper into a bowl. <EOS>', '<BOS> ...
4	dict	2	{'caption': ['<BOS> Someone is burning the tops of two cameras with a t ...
5	dict	2	{'caption': ['<BOS> Clouds are quickly floating past a building. <EOS>'] ...
6	dict	2	{'caption': ['<BOS> A older Jewish man slowly speaking. <EOS>', '<BOS> ...
7	dict	2	{'caption': ['<BOS> A woman peels an apple. <EOS>', '<BOS> A woman is p ...
8	dict	2	{'caption': ['<BOS> A man is performing yoga. <EOS>', '<BOS> A man does ...
9	dict	2	{'caption': ['<BOS> A man in a suit is doing one-handed push-ups. <EOS> ...
10	dict	2	{'caption': ['<BOS> Someone pours water into a kettle from a glass meas ...
11	dict	2	{'caption': ['<BOS> A phone is off the hook, and a bird appears to be d ...
12	dict	2	{'caption': ['<BOS> A large dog gets something out of a refrigerator. < ...
13	dict	2	{'caption': ['<BOS> A man is erasing a chalk board. <EOS>', '<BOS> One ...
14	dict	2	{'caption': ['<BOS> Baby pandas are crawling on a bench. <EOS>', '<BOS> ...

caption - Dictionary (2 elements)

caption - List (18 elements)

Index	Type	Size	Value
0	str	39	<BOS> A woman goes under a horse. <EOS>
1	str	61	<BOS> A woman crawls under a horse and gets a surprise. <EOS>
2	str	103	<BOS> A girl is jutting out her head from between the back legs of a h ...
3	str	48	<BOS> A horse defecated on a woman's head. <EOS>
4	str	122	<BOS> A horse excretes on the head of a woman as she gets in between t ...
5	str	52	<BOS> A horse is defecating on a woman's head. <EOS>
6	str	37	<BOS> A horse poops on a woman. <EOS>
7	str	37	<BOS> A horse poops on a woman. <EOS>
8	str	74	<BOS> A lady was pooped on because she was under a horse's tail end. < ...
9	str	131	<BOS> A woman crawls under a horse and when she sticks her head betwee ...
10	str	67	<BOS> A woman goes under a horse's behind and gets pooped on. <EOS>
11	str	44	<BOS> A woman goes underneath a horse. <EOS>
12	str	46	<BOS> A woman is crawling under a horse. <EOS>
13	str	52	<BOS> A woman is getting pooped on by a horse. <EOS>
14	str	50	<BOS> A woman is walking between horse legs. <EOS>
15	str	68	<BOS> A woman is walking under horse and it poops on her head. <EOS>
16	str	81	<BOS> As the woman went underneath the horse, the horse pooped on her ...
17	str	64	<BOS> The horse pooped on the lady who was underneath him. <EOS>

## 3. Build vocabulary:

word\_counts - Dictionary (8748 elements)

Key	Type	Size	Value
ad.	int	1	2
add	int	1	5
added	int	1	95
adding	int	1	228
adding,	int	1	2
addresses	int	1	2
adds	int	1	63
adds,	int	1	1

word\_counts - Dictionary (8748 elements)

Key	Type	Size	Value
swim	int	1	7
swimmer	int	1	2
swimmers	int	1	5
swimmers.	int	1	1
swimming	int	1	99
swimming.	int	1	17
swimmingpool	int	1	1
swims	int	1	11
swims.	int	1	4
swing	int	1	5
swing.	int	1	7
swinging	int	1	31

## 4. Training setup:

- Dataset = MSVD (1450/100, training/testing)
- Epochs = 30
- iteration = 2000
- Batch size = 10
- Learning rate = 0.0001

## 5. Training process tracking:

- Epoch 1: Training just started, weights are initializing, results tend to randomness.

```
Epoch: 0 iteration: 1000 , loss: 0.394902, time: 106.066846
Epoch: 1 iteration: 0 , loss: 0.349419, time: 105.669187
Epoch: 1 iter: 0 save succeeded!
.....
Output Caption:

A man is playing a guitar.
A man is slicing a bowl.
A man is slicing a piece of a bowl.
A woman is slicing a piece of a bowl.
A woman is slicing a piece of a bowl.
A man is playing a guitar.
A man is playing a guitar.
A man is playing a guitar.
A man is playing a guitar.
A woman is slicing a piece of a bowl.
.....
GT Caption:

<BOS> A gymnast is doing back flips then falls.
<BOS> The cat swatted someone from inside a box.
<BOS> A woman is putting a dish of liquid in a microwave.
<BOS> A woman is skewering shrimp.
<BOS> A man empties rice roni from a small carton box to a pan containing melted butter.
<BOS> A band is playing instruments.
<BOS> A small car explodes in a field.
<BOS> A man shoots another man on a horse.
<BOS> A boy is kicking a ball.
<BOS> The lady cut up shrimp and put it in the salad.
.....
```

Epoch: 1 niter: 0



- Epoch 2: The model started to generate some right sentences in the correct order.

Epoch: 3 iteration: 0 , loss: 0.488642, time: 106.471864  
Epoch: 3 iter: 0 save succeeded!

.....

Output Caption:

A woman is slicing a potato.  
A group of people are dancing.  
A dog is walking on a dog.  
A woman is riding a horse.  
A dog is eating a dog.  
A man is pouring a gun.  
A panda is eating a dog.  
A man is cutting a piece of a knife.  
A man is putting a gun.  
A woman is slicing a potato.

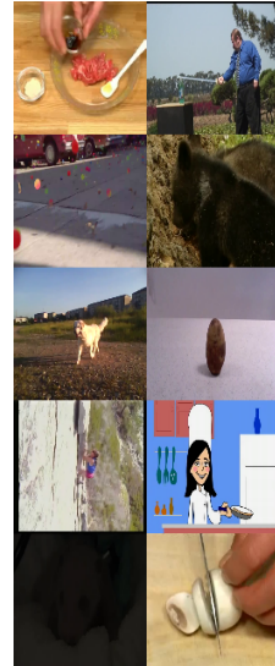
.....

GT Caption:

<BOS> The lady poured soy sauce and seasonings on the tomatoes.  
<BOS> A wall of bouncing balls descend a street.  
<BOS> A dog is waving its tail and barking.  
<BOS> A girl is climbing a rock.  
<BOS> Someone with gloved hands is listening to a baby panda through a stethoscope.  
<BOS> A man is cutting a bottle with a sword.  
<BOS> Two bears investigate a pile of dirt.  
<BOS> A large green ball is rolling into a potato.  
<BOS> The cartoon girl flipped a pancake in a frying pan.  
<BOS> A woman is cutting mushrooms.

.....

Epoch: 3 niter: 0



- Epoch 5: Output accuracy improves.

Epoch: 5 iteration: 0 , loss: 0.319838, time: 118.111180  
Epoch: 5 iter: 0 save succeeded!

.....

Output Caption:

A woman is putting ingredients in a bowl.  
A person is adding a piece of meat in a bowl.  
A woman is putting a piece of bread.  
A man is slicing a piece of bread.  
A man is riding a horse.  
A man is pouring oil into a pot.  
A man is pouring oil into a pot.  
A cat is playing with a ball.  
A woman is talking on a phone.  
A girl is playing a flute.

.....

GT Caption:

<BOS> A woman is cracking eggs.  
<BOS> A man is beating two eggs in a bowl with a fork.  
<BOS> A girl is cutting a piece of plastic.  
<BOS> A man is slicing a cucumber.  
<BOS> The boy hit a sign on his bike and fell off.  
<BOS> The man put water into the tomato sauce can.  
<BOS> A man is adding oil to a bowl.  
<BOS> A cat is eating out of a bowl and a puppy is biting the cats ear.  
<BOS> The man ate a banana.  
<BOS> A girl plays the violin.

.....

Epoch: 5 niter: 0



- Epoch 9: Loss decreased, the result is getting better.

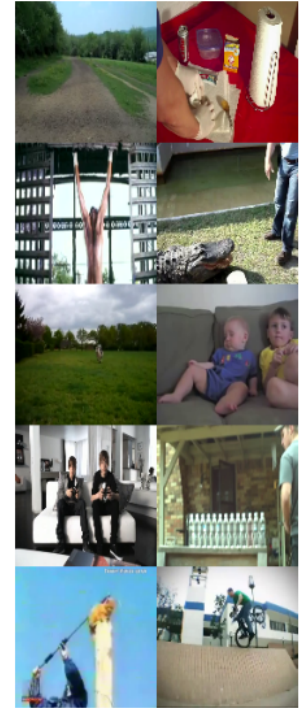
```
Epoch: 9 iteration: 0 , loss: 0.263892, time: 114.495699
Epoch: 9 iter: 0 save succeeded!
.....
Output Caption:
```

```
A woman is riding a horse.
A man is doing push ups.
A woman is riding a horse.
Two women are playing video games.
A cat is jumping into a car.
A man is cleaning a cd.
A man is cutting a dead deer.
A baby is falling asleep on a couch.
A man is cutting a piece of wood.
A man is riding a motorcycle.
```

```
.....
GT Caption:
```

```
<BOS> A woman is riding on a horse.
<BOS> The man did chin-up in the doorway.
<BOS> The rider trotted on the horse across the field.
<BOS> Two boys are playing video game.
<BOS> A cat jumps from a utility pole.
<BOS> A man is cleaning a DVD.
<BOS> The man pet the nose of the alligator.
<BOS> A boy is slipping while sitting on a couch.
<BOS> Using a knife, a man cuts through a row of bottles that are full of water.
<BOS> A man is performing stunts with his bike.
```

Epoch: 9 niter: 0



- Epoch 29: Training done, the result is very close to the ground truth.

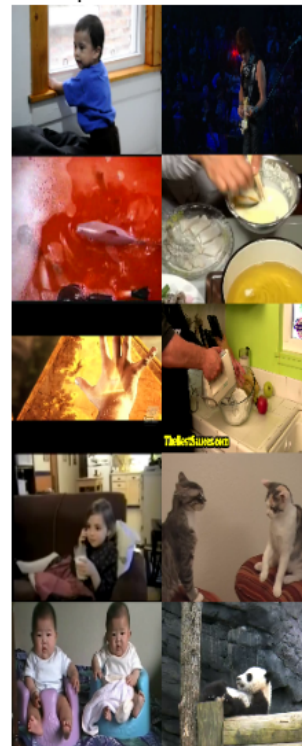
```
Epoch: 29 iteration: 0 , loss: 0.173709, time: 105.581422
Epoch: 29 iter: 0 save succeeded!
.....
Output Caption:
```

```
A boy is looking out a window.
A toy fish is moving in the water.
A man is staring at his hand.
Two girls are talking on the phone.
A baby is sneezing and other baby.
A man is playing the guitar.
A person is frying something.
A man is mixing ingredients in a bowl.
Two cats are fighting.
A panda is lying on logs.
```

```
.....
GT Caption:
```

```
<BOS> A small boy has placed his hands on the window pane and
<BOS> Bath toys are swimming.
<BOS> A bright triangle shape appears on the hand of a man.
<BOS> Two girls are talking in telephone.
<BOS> The babies are sitting down.
<BOS> A man is playing a guitar.
<BOS> The woman is frying vegetables.
<BOS> The man mixed the dough in a bowl.
<BOS> Two cats are playing while sitting on stools.
<BOS> A panda is lounging around.
```

Epoch: 29 niter: 0



6. Bleu evaluation:

- Score:  $0.61 > \text{baseline} = 0.6$

```
\MLDS_hw2_1_data>python bleu_eval.py test  
_output.txt  
Average bleu score is 0.6128513171038914
```

7. Summary:

- Total time cost: 6340 seconds  $\sim$  1.8 hours
- Loss plot:

