Waseda_Meisei_SoftBank at TRECVID 2022 Video to Text Description

Hiroki Takushima (presenter)
SoftBank Corporation
Kazuya Ueki
Meisei University, Waseda University
Takayuki Hori
SoftBank Corporation, Waseda University
Yuma Suzuki, Hideaki Okamoto
SoftBank Corporation

1. Overview
2. Methods
3. Experiments
4. Results
5. Discussion
6. Conclusion

Overview

Overview (Task)

- **VTT**（Video to Text Description）
  - Generating sentences from videos using natural language
  - Use temporal/spatial video features and audio

GenerateText：A man is playing frisbee a dog.

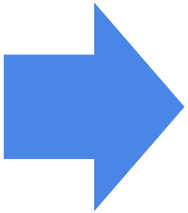
Overview（Competition）

- **TRECVID VTT 2022 regulation**
  - Generate English caption for 3~10 sec videos (and generate confidence score)
  - desire to consist of 4 content
    - Who : who is in the video? (people, animals, objects)
    - What : what are the objects and entities doing? (action or state)
    - Where : where was the video shot? (geographical or architectural location)
    - When : when was the video taken? (time of day, season, etc.)
  - Max 4 submissions each teams
  - select primary in all submissions

Methods

Strategy

- **Our Strategy**
  - Reduce redundancies in videos
  - TRECVID VTT 2022 Dataset audio is multilingual, so only video features are used
  - Use a pre-trained model because the training data is small

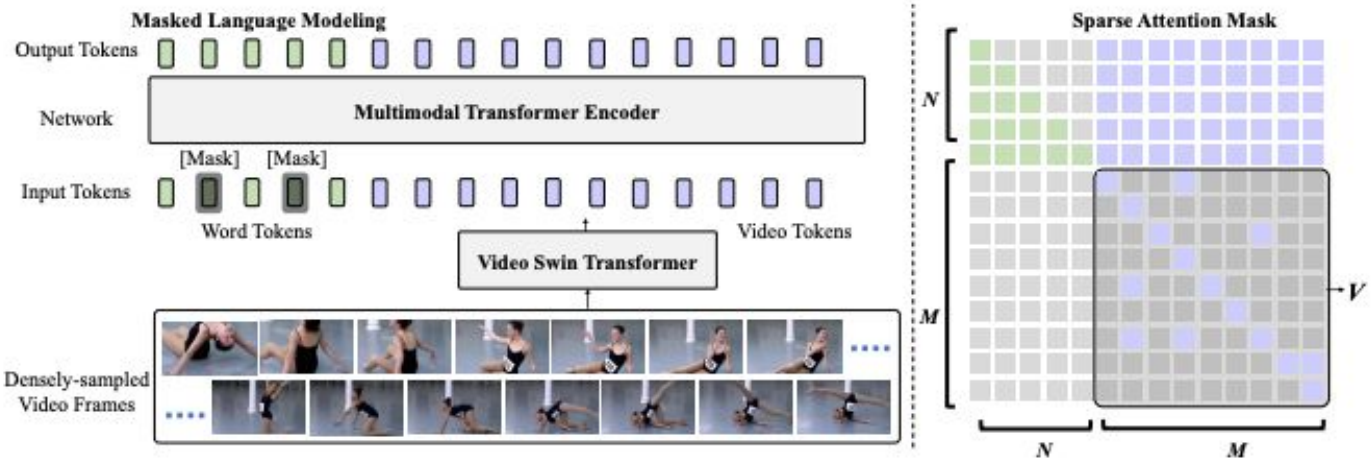


Use SwinBERT,[1] a SOTA model of VTT

[1]Lin, Kevin and Li, Linjie and Lin, Chung-Ching and Ahmed, Faisal and Gan, Zhe and Liu, Zicheng and Lu, Yumao and Wang, Lijuan SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning(CVPR2022)

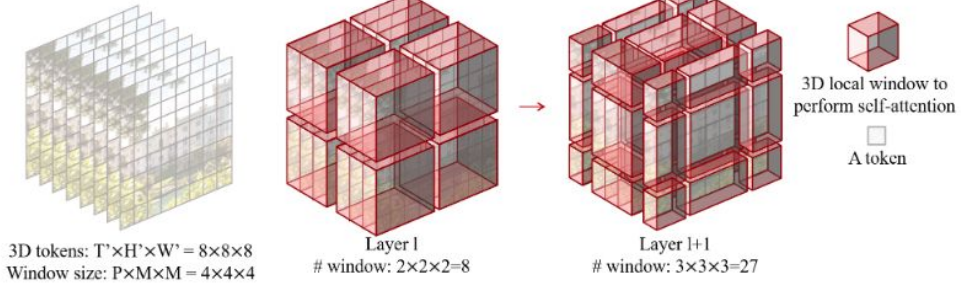
Methods - SwinBERT

- **SwinBERT**
  - End2End Video captioning model
    - **Encoder : Video Swin Transformer**
    - **Decoder : Multimodal Transformer Encoder**



Methods - SwinBERT

- **Encoder : Video Swin Transformer**
  - Visual feature extractor based on Transformer
  - **Patch-marge** : Split into N x N with patch like ViT
  - **SW-MSA** : Recognition between adjacent windows is possible by alternating windows
  - →Faster than sliding and still as accurate
  - **RelativePositionBias** : Adjust attention strength by relative position of patch



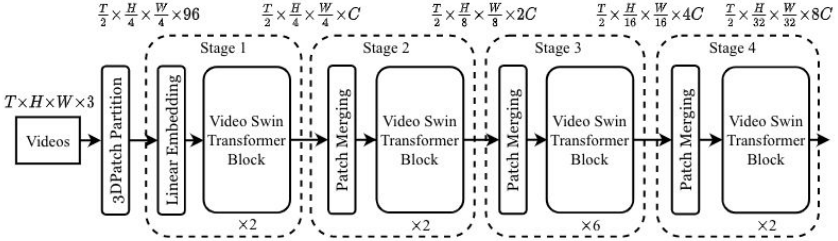
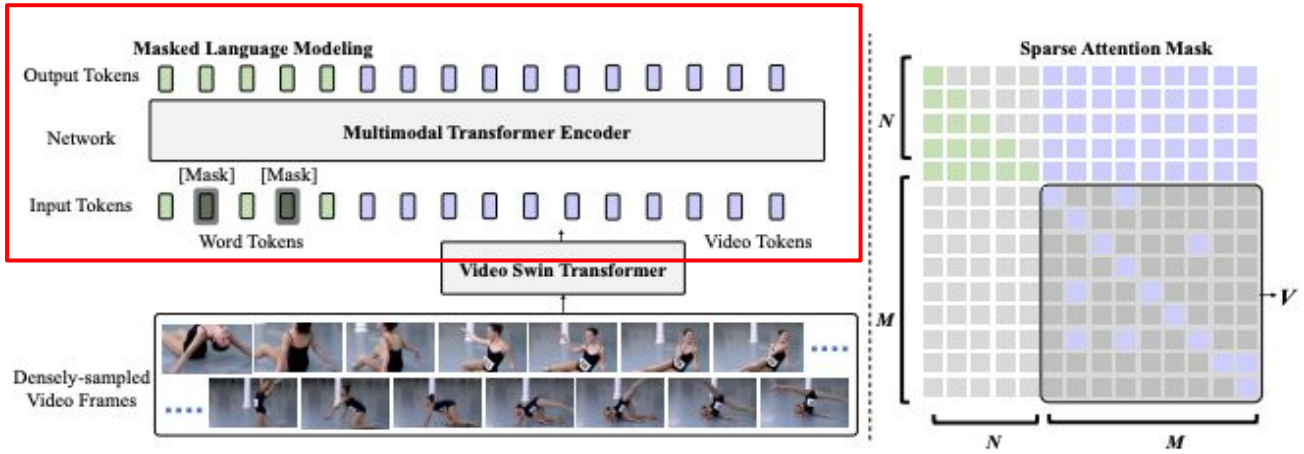


Figure 1: Overall architecture of Video Swin Transformer (tiny version, referred to as Swin-T).

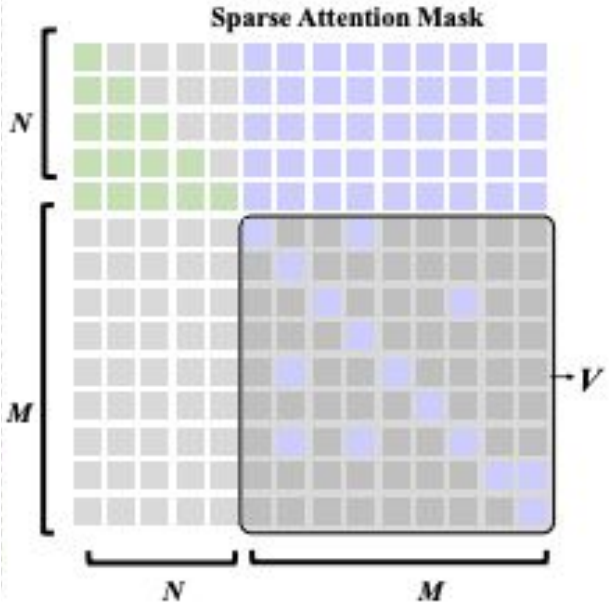
Methods - SwinBERT

- **Decoder : Multimodal Transformer Encoder**
  - Generate natural language description from textual and visual modality inputs
  - Training : Masked language model
  - Inference : Perform Seq2Seq generation



Methods - SwinBERT

- **Decoder : Multimodal Transformer Encoder**
  - **Learnable sparse Attention MASK**
    - **text→text**: can access only past tokes
    - **text→visual**: can access all
    - **visual→text**: can't access
    - **visual→visual**: train attention mask
      - **→reduce redundancies**



Experiments

Datesets

- **Dataset：V3C1**
  - Use V3C1 which consist of Vimeo Creative Commons Collection (V3C)
  - **Test data**
    - 2008 videos (+300 progress test)
    - 3～10sec per video
    - 5 captions per video
  - **Train / Develop data**
    - TRECVID data from 2016 years to 2021 years（10862 videos）
    - 2~5 captions per video

Experiments

- **Pretraining :**
  - Datasets : VATEX Dataset
  - AdamW optimizer : warmup 10% steps followed by linear decay
  - Use max 32 frames
- **Finetuning :**
  - Datasets : V3C1 Dataset
  - Don’t train leanable attention mask

Evaluation Metrics

- **Metrics**
  - **Automatic metrics**
    - METEOR, BLEU, CIDEr, CIDEr-D, SPICE
  - **Direct Assessment(DA)**
    - Evaluate primary runfile only by cloud worker →**TBD**
  - **Semantic similarity metric(STS)**
    - Measuring the semantic relevance of features →**Add Appendix**

Results

Results

- Submit runfiles Fintunig TRECVID-VTT
- For primary, check the generated sentences and select

Runfile	CIDER	CIDER-D	BLEU	METEOR	SPICE
1	**0.415**	0.178	0.033	0.260	0.077
2	0.348	0.141	0.026	0.252	0.084
3	0.350	0.150	0.028	0.260	0.087
4	0.388	**0.182**	**0.037**	**0.286**	**0.100**

※Best runfile results

Results

- Max 3rd, Average rank is 4.2

	CIDER	CIDER-D	BLEU	METEOR	SPICE
ELT_01	4th	2nd	4th	5th	3rd
VIDION	3rd	6th	6th	6th	6th
kslab	2nd	3rd	2nd	4th	5th
MLVC_HDU	6th	5th	3rd	2nd	2nd
RUCAIM3-Tencent	1st	1st	1st	1st	1st
WasedaMeiseiSoftbank	5th	4th	5th	3rd	4th

※Best runfile results

Discussion

Output samples

GT1:	A person surfing on the frothy white waves in the ocean on a sunny day.
GT2:	A man surfs in the ocean going up and down in the waves with his surfboard on a cloudy day.
GT3:	A surfer is jumping over waves on a cloudy day
GT4:	A man is water surfing on a large wave in the daytime.
GT5:	A guy is surfing in big and high waves in the sea or ocean in the daytime.

Run1:	a man is surfing on a wave and then falls off.
Run2:	a surfer rides a surfboard surfs through the water on a sunny day.
Run3:	a man on a surfboard rides a wave and jumps off his head.
Run4:	a man rides a surf board through a ocean on a sunny day.

Output samples

GT1:	A group of people are on top of a windy mountain looking down on a city as one of the men is showing them something.
GT2:	One man out of four people standing on top of a mountain and watching down is explaining something to them using his hands in the daytime.
GT3:	A group of people sight seeing on a mountain top look over a town.
GT4:	Three men and a long haired woman are standing on a hill overlooking a settlement in the valley below them on a windy day.
GT5:	Four white people standing on a hill overlooking a small town situated in a valley on an overcast day.

Run1:	a group of people are standing on a cliff and one of them is standing on the ground
Run2:	a young woman with a black shirt and a black shirt and a black shirt and a black
Run3:	a man in a white shirt and a woman are taking a selfie on a rocky mountain
Run4:	a young woman walks towards a man on a rocky day as a young woman walks up to

V3C1 train data samples

Users > takushima > workspace > trecvid > vtt_ground_truth.txt

```
1 a man sings in a car
1 man shakes his body in a car at daytime
2 in the daytime, two sportsmen walk in the sports field, while an other man moves on the wheel
2 on a baseball field a person in dark clothes is moving from right to left without walking
3 a basketball player wearing number 30 shirt is scoring a goal on a basketball court against a red wearing team
3 basketball player scoring a three pointer on a basketball field
4 people are going downstairs on a stage
4 several people walking down a staircase on a stage
5 2 women shake their bottoms
5 two women try to dance in a room.
6 a group of people riding a free fall tower
6 a roller coaster moves upwards at daytime
7 a man is scared by a picture of a boy on a mirror in a bathroom
7 white male taking a selfie in a bathroom mirror
8 2 men on a stage hug and walk away
8 two young man are standing on a stage embracing each other and one claps the other one's ass.
9 3 men are hugging each other
9 a man performs in movies
10 a man speaks into a microphone indoors
10 white man with glasses giving an interview in the interview area
11 a basketball player walks through a group of people
11 a big man in yellow cloths passes by some chairs in a large hall.
12 in the daytime, a cat walks at home and meows and another cat approaches to it
12 a little cat is following the camera and meows to it
13 2 men train in a gym
13 2 men run in a gym
```

- **w/o period** : Variations of sentences with and without periods
- **number** : Numeric and alphanumeric variations
- **human** : Many descriptions of what you are wearing
- **color info** : Many descriptions of color infomation

Discussion

- **w/ or w/o period ←critical to our approach**

  VATEX dataset has periods

  V3C1 dataset with and without periods

  → Period output becomes ambiguous in Finetuning for pre-trained in VATEX

  → Impression that the model does not know the end of the sentence

- **Difference in number description**

  V3C1 Train data has both alphanumeric and numeric descriptions

  V3C1 Test data has description only in alphanumeric characters

  → Deviation occurs in the output

Discussion

- **Descriptions about people**

  V3C1 data has many descriptions of information that people wear

  The point is whether this is described

- **Description about color**

  V3C1 data has many descriptions about colors

  In particular, there are many descriptions about what the person above is wearing

Conclusion

Conclusion

- SwinBERT reduces redundancy for videos and can effectively generate even less data by pre-training model
- Finetuning SwinBERT with V3C1 dataset
- From the generated results, we found the features and issues of the V3C1 dataset
- As a result, the highest ranking was 3rd, and the average rank was 4.2.

Appendix

Datesets

- **Dataset：V3C1**
  - Use V3C1 which consist of Vimeo Creative Commons Collection (V3C)
  - Test data
    - 2008 videos (+300 progress test)
    - 3～10sec per video
    - 5 captions per video
  - Train / Develop data
    - TRECVID data from 2016 years to 2021 years（10862 videos）
    - 2~5 captions per video
    - data details
      - Videos with IDs ~6475 videos from Twitter Vine.
      - Videos with IDs 6476 - 7485 are from our Flickr dataset.
      - Videos with IDs 7486~ are from the V3C dataset.
      - Videos with IDs from 1 to 3528 have between 2 to 5 captions.
      - Videos from 3529 onwards have 5 captions each.

Result

- CIDER

	Run1	Run2	Run3	Run4	Best	Rank
ELT_01	0.507	0.103	0.243	0.234	0.507	4th
VIDION	0.595	0.589	0.607	0.611	0.611	3rd
kslab	0.619	0.163	0.510	0.141	0.619	2nd
MLVC_HDU	0.361	0.346	0.361	0.361	0.361	6th
RUCAIM3-Tencent	0.940	0.936	0.936	0.942	0.942	1st
WasedaMeiseiSoftbank	0.415	0.348	0.350	0.388	0.415	5th

Result

- CIDER-D

	Run1	Run2	Run3	Run4	Best	Rank
ELT_01	0.226	0.045	0.076	0.105	0.226	2nd
VIDION	0.098	0.099	0.108	0.113	0.113	6th
kslab	0.194	0.048	0.110	0.027	0.194	3rd
MLVC_HDU	0.179	0.166	0.179	0.179	0.179	5th
RUCAIM3-Tencent	0.594	0.575	0.602	0.592	0.602	1st
WasedaMeiseiSoftbank	0.178	0.141	0.150	0.182	0.182	4th

Result

- **BLEU**

	Run1	Run2	Run3	Run4	Best	Rank
ELT_01	0.069	0.012	0.014	0.034	0.069	4ht
VIDION	0.024	0.025	0.029	0.030	0.030	6th
kslab	0.081	0.0260	0.047	0.011	0.081	2nd
MLVC_HDU	0.0716	0.062	0.0716	0.0716	0.0716	3rd
RUCAIM3-Tencent	0.1350	0.131	0.1352	0.1353	0.1353	1st
WasedaMeiseiSoftbank	0.033	0.0263	0.028	0.037	0.037	5th

Result

- **METEOR**

	Run1	Run2	Run3	Run4	Best	Rank
ELT_01	0.248	0.178	0.169	0.194	0.248	5th
VIDION	0.212	0.211	0.220	0.221	0.221	6th
kslab	0.281	0.204	0.226	0.170	0.281	4th
MLVC_HDU	0.289	0.280	0.289	0.289	0.289	2nd
RUCAIM3-Tencent	0.412	0.409	0.414	0.413	0.414	1st
WasedaMeiseiSoftbank	0.2604	0.252	0.2603	0.286	0.286	3rd

Result

- **SPICE**

	Run1	Run2	Run3	Run4	Best	Rank
ELT_01	0.102	0.043	0.062	0.064	0.102	3rd
VIDION	0.073	0.073	0.077	0.077	0.077	6th
kslab	0.097	0.049	0.071	0.036	0.097	5th
MLVC_HDU	0.107	0.098	0.107	0.107	0.107	2nd
RUCAIM3-Tencent	0.182	0.180	0.184	0.183	0.184	1st
WasedaMeiseiSoftbank	0.077	0.084	0.087	0.100	0.100	4th

Result

- **STS（Semantic smilarity metrics）：Run1**

	TXT1	TXT2	TXT3	TXT4	TXT5
ELT_01	0.4211	0.4189	0.4199	0.4191	0.4151
VIDION	0.3966	0.3897	0.3938	0.3908	0.3902
kslab	0.4194	0.4126	0.4126	0.4137	0.4177
MLVC_HDU	0.4176	0.3949	0.3928	0.3988	0.4188
RUCAIM3-Tencent	0.5380	0.5140	0.5175	0.5116	0.5349
WasedaMeiseiSoftbank	0.3563	0.3658	0.3669	0.3654	0.3564

Result

- **STS（Semantic smilarity metrics）：Run2**

	TXT1	TXT2	TXT3	TXT4	TXT5
ELT_01	0.2357	0.2401	0.2302	0.2378	0.2386
VIDION	0.3984	0.3867	0.3927	0.3930	0.3888
kslab	0.2705	0.2616	0.2603	0.2648	0.2703
MLVC_HDU	0.4087	0.3864	0.3822	0.3868	0.4061
RUCAIM3-Tencent	0.5327	0.5121	0.5102	0.5100	0.5332
WasedaMeiseiSoftbank	0.3849	0.3755	0.3616	0.3715	0.3843

Result

- **STS（Semantic smilarity metrics）：Run3**

	TXT1	TXT2	TXT3	TXT4	TXT5
ELT_01	0.3570	0.3413	0.3351	0.3361	0.3627
VIDION	0.4064	0.3947	0.3996	0.3985	0.3954
kslab	0.3690	0.3641	0.3629	0.3602	0.3688
MLVC_HDU	0.4176	0.3949	0.3928	0.3988	0.4188
RUCAIM3-Tencent	0.5386	0.5154	0.5151	0.5147	0.5379
WasedaMeiseiSoftbank	0.3881	0.3745	0.3689	0.3698	0.3856

Result

- STS（Semantic smilarity metrics）：Run4

	TXT1	TXT2	TXT3	TXT4	TXT5
ELT_01	0.3083	0.3034	0.2994	0.3034	0.3073
VIDION	0.4061	0.3951	0.3998	0.3984	0.3952
kslab	0.2535	0.2504	0.2460	0.2456	0.2539
MLVC_HDU	0.4176	0.3949	0.3928	0.3988	0.4188
RUCAIM3-Tencent	0.5351	0.5146	0.5131	0.5125	0.5333
WasedaMeiseiSoftbank	0.4148	0.4094	0.4023	0.4035	0.4184