LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts

Shuming Ma[1] Lei Cui[2] Damai Dai[1] Furu Wei[2] Xu Sun[1]

[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Microsoft Research Asia

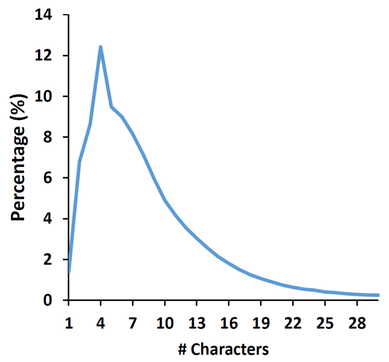
{shumingma, daidamai, xusun}@pku.edu.cn

{lecu, fuwei}@pku.edu.cn

The Live Video Comment Task and Dataset

Live video commenting, which is also called ''video barrage'' (''弹幕'' in Chinese or ''Danmaku'' in Japanese), is an emerging feature on online video sites that allows real-time comments from viewers to fly across the screen like bullets or roll at the right side of the screen.

Statistic	Train	Test	Dev	Total
#Video	2,161	100	100	2,361
#Comment	818,905	42,405	34,609	895,929
#Word	4,418,601	248,399	193,246	4,860,246
Avg. Words	5.39	5.85	5.58	5.42
Duration (hrs)	103.81	5.02	5.01	113.84



Interval	Edit Distance	TF-IDF	Human
0-1s	11.74	0.048	4.3
1-3s	11.79	0.033	4.1
3-5s	12.05	0.028	3.9
5-10s	12.42	0.025	3.1
>10s	12.26	0.015	2.2

The dataset is crawled from a popular video streaming website called BiliBili.com.

An Example from the Live Commenting Dataset

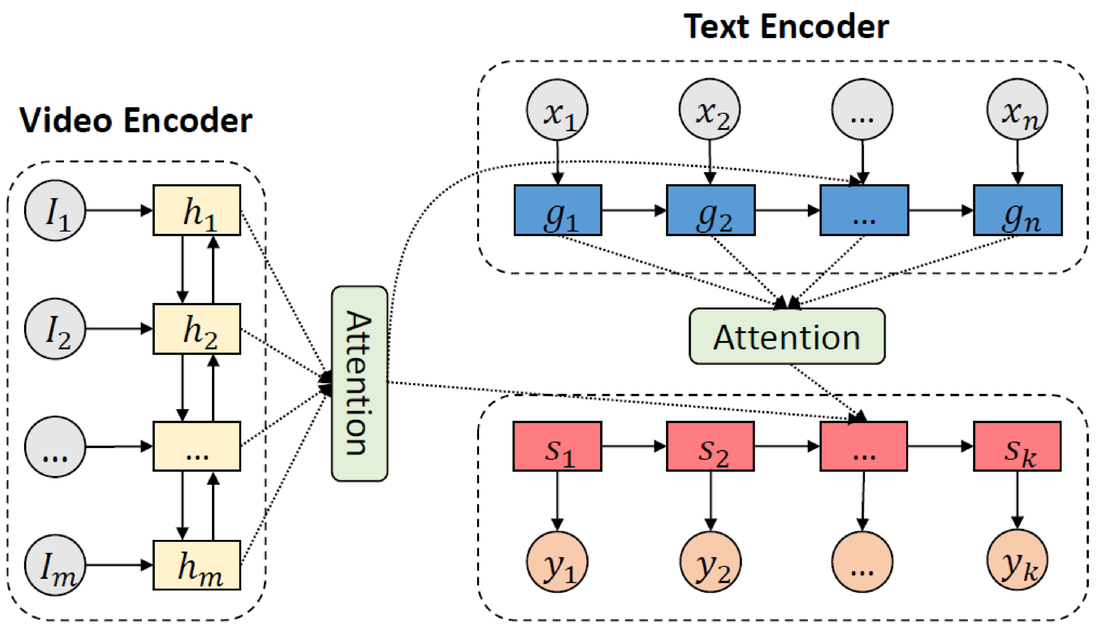
(a) 0:48 (b) 1:52 (c) 3:41

Three selected frames from the videos to demonstrate the content

Time Stamp	Comments
0:48	橙猫是短腿吗 (Is the orange cat short leg?)
1:06	根本停不下来 (Simply can't stop)
1:09	哎呀好可爱啊 (Oh so cute)
1:52	天哪这么多，天堂啊 (OMG, so many kittens, what a paradise!)
1:56	这么多只猫 (So many kittens!)
2:39	我在想猫薄荷对老虎也有用吗 (I am wondering whether the catmint works for the tiger.)
2:41	猫薄荷对老虎也有用 (Catmint also works for the tiger.)
3:41	活得不如猫 (The cat lives even better than me)
3:43	两个猫头挤在一起超可爱 (It's so cute that two heads are together)

Several selected live comments paired with the time stamps when the comments appear on the screen.

Live Commenting Models



Two approaches to generate the comments based on the visual contexts and the textual contexts.

The two approaches are based on two popular architectures for text generation: recurrent neural network (RNN) and transformer.

Both two models consist of a video encoder, a text encoder, and a comment decoder. We denote two approaches as Fusional RNN Model and Unified Transformer Model, respectively.

Experiments

	Model	#I	#C	Recall@1	Recall@5	Recall@10	MR	MRR
Video Only	S2S-I	5	0	4.69	19.93	36.46	21.60	0.1451
	S2S-IC	5	0	5.49	20.71	38.35	20.15	0.1556
	Fusional RNN	5	0	10.05	31.15	48.12	19.53	0.2217
	Unified Transformer	5	0	**11.40**	**32.62**	**50.47**	**18.12**	**0.2311**
Comment Only	S2S-C	0	5	9.12	28.05	44.26	19.76	0.2013
	S2S-IC	0	5	10.45	30.91	46.84	18.06	0.2194
	Fusional RNN	0	5	13.15	**34.71**	**52.10**	17.51	0.2487
	Unified Transformer	0	5	**13.95**	34.57	51.57	**17.01**	**0.2513**
Both	S2S-IC	5	5	12.89	33.78	50.29	17.05	0.2454
	Fusional RNN	5	5	17.25	37.96	**56.10**	16.14	0.2710
	Unified Transformer	5	5	**18.01**	**38.12**	55.78	**16.01**	**0.2753**

Model	Fluency	Relevance	Correctness
S2S-IC	4.07	2.23	2.91
Fusion	**4.45**	2.95	3.34
Transformer	4.31	**3.07**	**3.45**
Human	4.82	3.31	4.11

Our models outperform the existing video-to-text models in terms of both automatic evaluation and human evaluation.

Conclusions and Future Work

- **A new task: live video commenting**
- **A new dataset: the live comment dataset**
- **Two new baselines: Fusional RNN and Unified Transformer**
- **An evaluation protocol**

Future

- **A better video/image representation**
- **Generating low-frequent/surprising comments**
- **Integrating audio information**
- **Other tasks: video MT, video summarization, video QA**

Scan the QR code for the codes and dataset