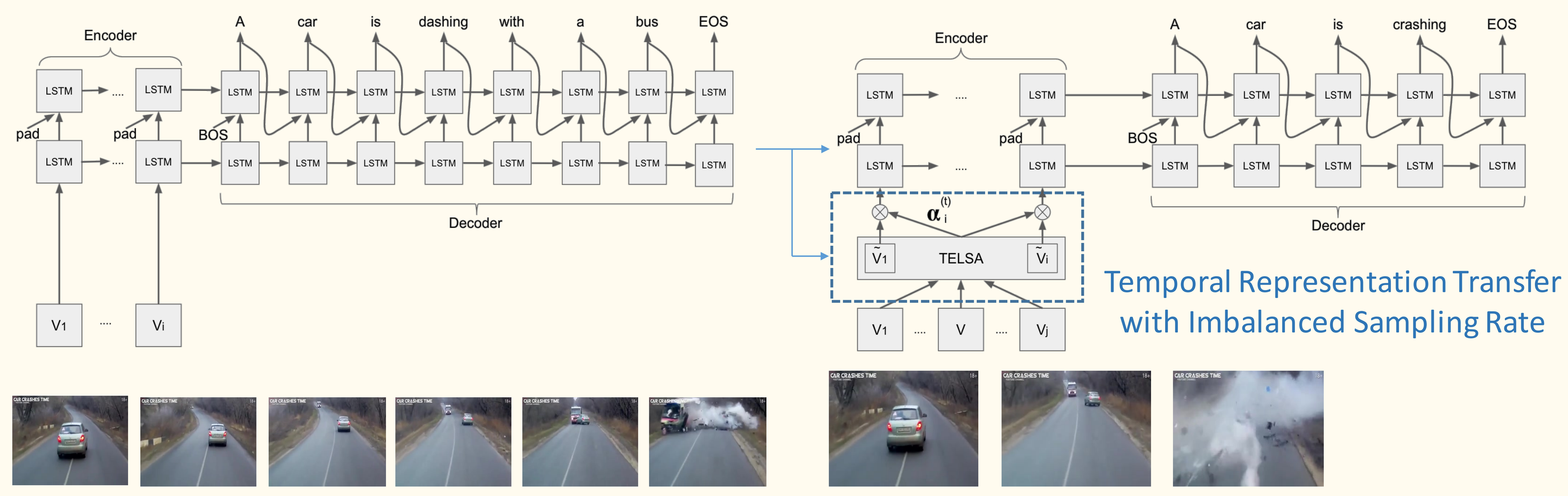
HOW SAMPLING RATE AFFECTS CROSS-DOMAIN TRANSFER LEARNING FOR VIDEO DESCRIPTION

Yu-Sheng Chou[1], Pai-Heng Hsiao[2], Shou-De Lin[1] and Hong-Yuan Mark Liao[3]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan
[2]Memorence A.I., Taipei City, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan

Two Issues Associated with Video-to-language Transfer Learning Problem

- **How to transfer knowledge learned from a more general dataset to a specific application domain dataset?**
- **How to generate stable video description results under different sampling rates?**
- We leverage a stacked LSTM encoder-decoder structure and propose a temporal embedding method to better retain temporal representation under different video sampling rates for the transfer learning task.

Architecture for Temporal Representation Transfer



- TELSA mechanism can transfer temporal embeddings and adjust visual representations in encoding phase (TELSA mechanism in dash line rectangle is only activated when fine-tuning on target domain).

Examples of Automatic Video Description at Different Sampling Rates on MSR-VTT

(a) set of frames grabbed by sparser sampling rate: "a car is crashing."

(b) set of frames grabbed by denser sampling rate: "a car is dashing with a bus."

Problem Formulation and Transfer Learning on Temporal Representation

- **Video-to-language translation with LSTM:**

$$p(y|x) = \prod_{t=1}^{m} p(y_t|h_{n+t-1}, y_{n+t-1}, z)$$

- **Temporal embedding learning with soft-attention (TELSA):**

$$z = \prod_{t=1}^{n} p(y_t|h_{t-1}, \alpha_i^{(t)} \tilde{v}_i)$$

$$\tilde{v}_i = E(v_i)$$

- **Dynamic weights $\alpha_i^{(t)}$:**

$$e_i^{(t)} = w^T tanh(W_a h_{t-1} + \tilde{v}_i + b_a)$$

$$\alpha_i^{(t)} = \frac{exp\{e_i^{(t)}\}}{\sum_{j=1}^{n} exp\{e_j^{(t)}\}}$$

Experiments

- **Single-domain Analysis: MSR-VTT.** We dealt with the imbalanced sampling rate problem within single domain, i.e., the trained source domain and the target domain were both in MSR-VTT.

Train:Test samples	METEOR	BLEU @1	@2	@3	@4
80:80	26.10	75.50	60.30	46.70	34.80
40:80	25.40	75.40	58.80	44.20	32.20
40:80+TELSA	**26.00**	**77.40**	**61.50**	**47.30**	**34.60**

- **Cross-domain Analysis: MSR-VTT to MSVD.** We handled the imbalanced sampling rate problem within cross-domain environment.

Method	Single Domain Source	Single Domain Target	Transfer Learning Fine-tuning	Transfer Learning ours
A: 40S:80T				
METEOR	26.80	26.70	28.14	**29.19**
BLEU@1	67.90	69.90	72.35	**74.49**
BLEU@2	50.30	54.10	57.25	**59.78**
BLEU@3	38.20	43.70	46.79	**49.26**
BLEU@4	27.10	33.20	36.81	**39.01**
B: 80S:80T				
METEOR	26.90	26.70	27.99	**28.55**
BLEU@1	69.10	69.90	72.38	**73.07**
BLEU@2	52.40	54.10	57.02	**57.49**
BLEU@3	40.60	43.70	46.45	**47.01**
BLEU@4	29.50	33.20	36.53	**36.67**
C: 120S:80T				
METEOR	26.00	26.70	27.30	**28.00**
BLEU@1	67.70	69.90	71.15	**72.08**
BLEU@2	50.60	54.10	56.10	**56.65**
BLEU@3	39.10	43.70	45.65	**46.13**
BLEU@4	27.60	33.20	35.30	**35.75**

Test Video Clip

Ground Truth: a man scores a goal in soccer.

	A: 40S:80T	B: 80S:80T	C: 120S:80T
Source Data Only:	a man is playing a video game.	a man is playing a football game.	a man is playing a football game.
Fine-tuning:	a man is playing football.	a man is playing football.	a football player is running down the field.
Fine-tuning + TELSA:	**a soccer player is playing the goal.**	**a man is playing a soccer game.**	a football player is playing football.