CAPTURING LONG-RANGE DEPENDENCIES IN VIDEO CAPTIONING

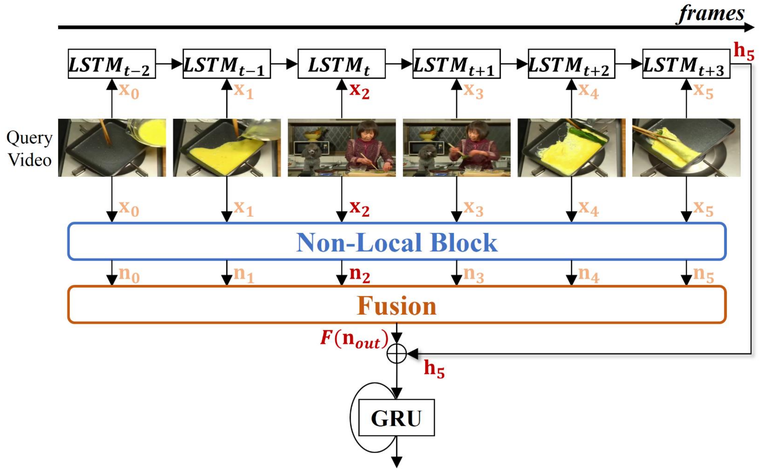
Jaeyoung Lee[1,3], Yekang Lee[1], Sihyeon Seong[1,3], Kyungsu Kim[2], Sungjin Kim[2], and Junmo Kim[1,3]

[1]Korea Advanced Institute of Science and Technology, [2]Samsung Research, [3]Mofl, Inc.

INTRODUCTION

- Video captioning requires both video understanding and natural language processing.
- Most video captioning models rely on recurrent models including LSTM.
- These model still suffer from long-range dependency problem in video captioning.

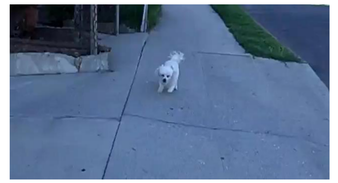
METHOD



- Given video $V = (v_0, v_1, \dots, v_n)$, we sampled k frames for all videos and extracted the features $X = (x_0, x_1, \dots, x_k)$ using ResNet-50 or ResNet-152.
- The extracted features are fed into the LSTM cell to make a context vector h_k.
- The extracted features are also fed into the non-local block[1].
- The outputs from the non-local block are collected and linearly embedded to form $F(n_{out})$.
- GRU uses a combination of h_k and $F(n_{out})$ to generate words sequentially.

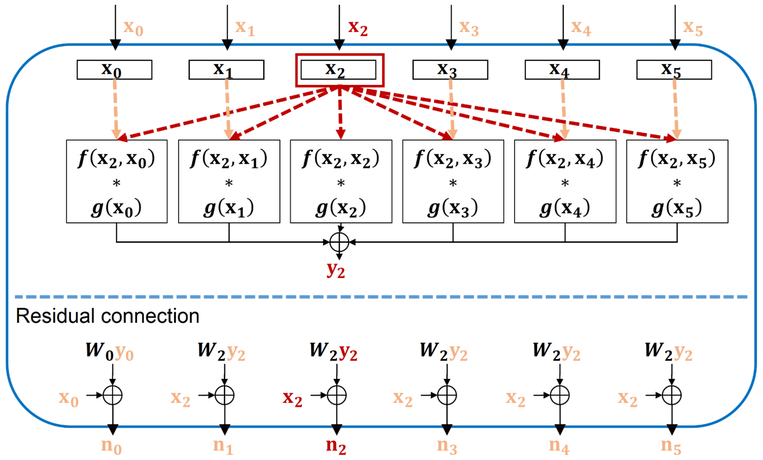
RESULTS

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
S2VT [2]	-	-	-	-	29.8	-	-
Y. Liu *et al.* [18]	80.2	69.0	60.1	51.1	32.6	-	-
h-RNN [3]	81.5	**70.4**	**60.4**	49.9	32.6	-	65.8
Attention fusion [19]	-	-	-	**52.4**	32.0	-	68.8
BA encoder [4]	-	-	-	42.5	32.4	-	63.5
PickNet [20]	-	-	-	52.3	33.3	69.6	76.5
Baseline	76.8	64.2	54.7	44.8	32.5	68.8	75.0
J. Lee *et al.* [9]	78.3	65.8	56.1	46.1	33.0	69.5	78.3
Ours (ResNet-50)	78.9	66.8	56.9	46.6	33.4	69.9	81.0
Ours (ResNet-152)	**82.9**	69.9	59.8	49.7	**33.7**	**71.7**	**84.5**



GT: a dog is running down the sidewalk.
Baseline: a dog is walking.
J. Lee et. al.: a dog is walking on the ground.
Ours: a dog is walking on the sidewalk.

NON-LOCAL BLOCK[1]

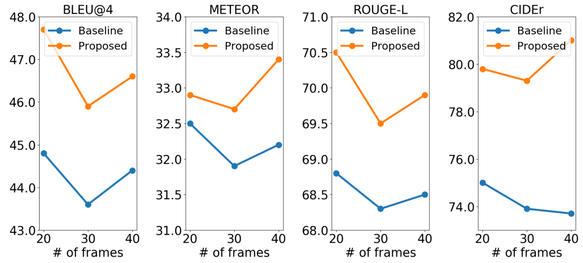


$y_i = \frac{1}{C}\sum_{\forall j} f(x_i, x_j) g(x_j)$, where C is a normalization factor.

Pairwise function f can be $f(x_i, x_j) = e^{x_i W_\theta^T W_\phi x_j}$.

$n_i = W_z y_i + x_i$ (residual connection).

PERFORMANCES/FRAMES



CONCLUSION

- A non-local block can complement a LSTM cell in terms of temporal capacity in video captioning.
- A non-local block encourages the network to utilize less trivial and more informative words.

[REFERENCES]
1. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.