INTRODUCTION

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

METHOD

• Given video \( V = (v_0, v_1, \ldots, v_n) \), we sampled \( k \) frames for all videos and extracted the features \( X = (x_0, x_1, \ldots, x_k) \) 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

\[
y_i = \frac{1}{C} \sum_{j \neq i} f(x_i, x_j) g(x_j), \text{ where } C \text{ is a normalization factor.}
\]

Pairwise function \( f \) can be \( f(x_i, x_j) = e^{x_i W_x^f x_j} \).
\( n_j = W_x y_j + x_i \) (residual connection).

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]