Deep Learning-Based Video Captioning in Bengali

Supervisor

Faisal Muhammad Shah
Associate Professor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Group Members

Saiful Islam	16.01.04.040
Amir Hossain Raj	16.01.04.056
Aurpan Dash	16.01.04.065
Ashek Seum	16.01.04.067

Presentation Outline

01 Introduction

02 Literature Review

03 Proposed Methodology

04 Results and Evaluation

05 Limitations and Future Works

01 Introduction

Video Captioning

- Generation of natural language phrases explaining the contents of video frames.

Caption: A man puts a container into microwave and activates it.

Video Captioning in Bengali

- Generation of phrases explaining the contents in Bengali language.

Caption: একজন ব্যক্তি একটি মাইক্রোওয়েভে একটি থালা রাখেন এবং এটি শুরু করেন

Motivation

- Navigation for visually impaired people
- Sign-language to natural-language conversion
- Real time suspicious activity detection
- Better Human-Robot interaction
- Storage minimization

Objective

- Building a model to extract visual features from videos and generate natural language captions in Bengali.
- Select adequate dataset depending on the videos on a variety of activities.
- Learn from available works related to this field
- Collecting some state-of-the-art methods for our Bengali caption generation model.
    - object detection,
    - spatio-temporal feature extraction,
    - language generation tasks.
- Trying out different combinations to face the challenge of video captioning in Bengali.

Thesis Contribution

- Translated all the captions of Microsoft Video Definition (MSVD) dataset to Bengali using Google Translate API.
- Removed the irrelevant translations and some of the rare words.
- Developed an encoder-decoder based model which can successfully generate captions in Bengali from input videos.

02 Literature Review

Research Papers Regarding Video Captioning in English

Ref.	Year	Title	Video Captioning technique and procedures
[1]	2018	Reconstruction network for video captioning	• Inception-V4 used as the encoder • LSTM+GRU used for decoder part • Backward Flow done through NMT mechanism and image segmentation.
[2]	2019	Joint event detection and description in continuous video streams	• (C3D) architecture employed as encoder • SPN predicts the activity proposals' duration. • Two level of LSTM is used.
[3]	2019	Hierarchical vision-language alignment for video captioning	• GoogLeNet with Batch Normalization • three parallel encoder-decoder streams • attention-based encoder and an alignment-embedded decoder

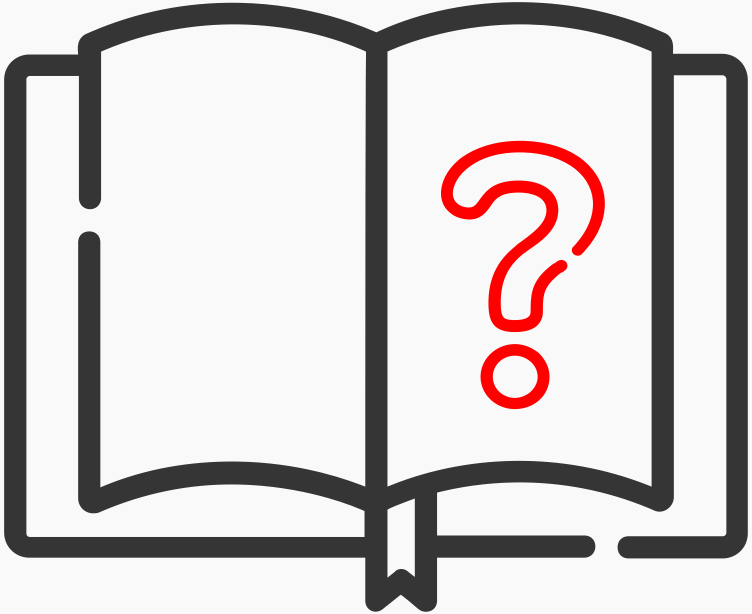
[1] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7622–7631, 2018.
[2] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, "Joint event detection and description in continuous video streams," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 396–405, IEEE, 2019.
[3] J. Zhang and Y. Peng, "Hierarchical vision-language alignment for video captioning," in International Conference on Multimedia Modeling, pp. 42–54, Springer, 2019.

Research Papers Regarding Image Captioning in Bengali

Ref.	Year	Title	Video Captioning technique and procedures
[4]	2019	Chitron: An automatic bangla image captioning system	• trained on 15,700 images and, 300 images are considered as the test data • two inputs: image, and sequence of tokens. • VGG16 model used as the pre-trained image model. • Stacked LSTM layers as one-word-at-a-time strategy to predict caption
[5]	2019	Oboyob: A sequential-semantic bengali image captioning engine	• a Bengali rule based stemmer has been used • Pre-trained Inception-ResNet and VGG-16 models used for images' feature extraction. • FastText library's models utilized for pre-trained word embedding. • introduced a pre-compiled word embedding model.

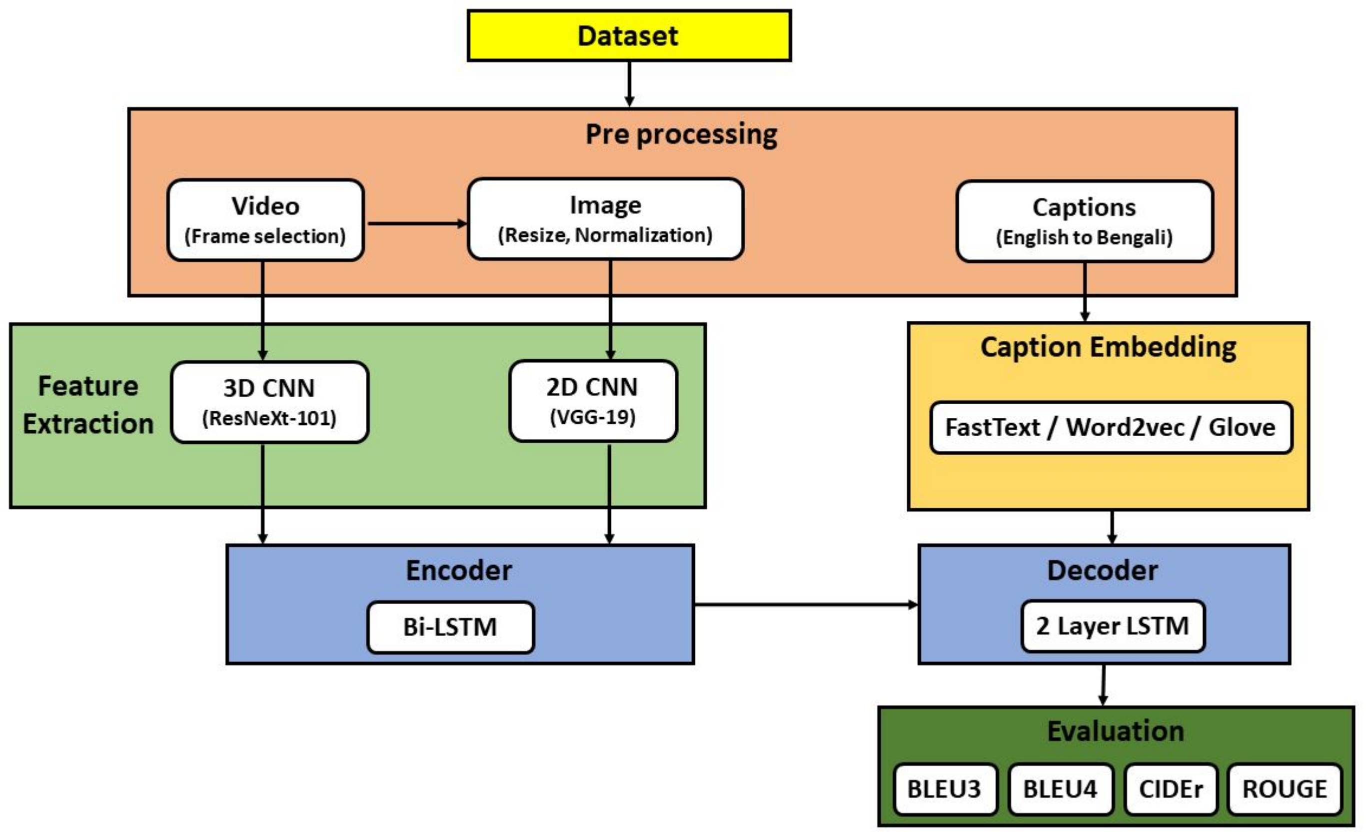
[4] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, "Chittron: An automatic bangla image captioning system," Procedia Computer Science, vol. 154, pp. 636–642, 2019.
[5] T. Deb, M. Z. A. Ali, S. Bhowmik, A. Firoze, S. S. Ahmed, M. A. Tahmeed, N. Rahman, and R. M. Rahman, "Oboyob: A sequential-semantic bengali image captioning engine," Journal of Intelligent & Fuzzy Systems, no. Preprint, pp. 1–13, 2019.

Research Papers Regarding Video Captioning in Bengali



03 Proposed Methodology

Main Architecture of Our Model



Dataset Description

Microsoft Video Description Corpus (MSVD)

- 1,970 single event video clips
- 10 to 25 seconds
- 85,550 English captions, 43 captions per video
- Vocabulary contains 13,010 English unique words

Modifications:

- **85,550 translated Bengali captions**
- **Vocabulary contains 7,105 Bengali unique words with 4 tokens**

Pre-processing

1. Video Pre-processing

- Selected 3 frames out of 30 frames per second
  (one frame from every 10 frames)
- Gathered exactly 32 frames per video-clips of variable lengths through replication and truncation

Pre-processing (contd.)

2. Image Pre-processing

- Resized to the dimension of 224 x 224 x 3
- Normalized images using means and std. daviations
  - transformed pixel values into a range of [0, 1]
  - mean = [R: 0.485, G: 0.456, B: 0.406]
  - std. daviations = [R: 0.229, G: 0.224, B: 0.225]

Pre-processing (contd.)

3. Caption Pre-processing

- Translated English captions to Bengali using Google Translation API
- Removed unwanted noise from translations
- sentences were tokenized to create vocabulary of unique words
- <start>, <end> tokens added to mark beginning and end of sentence
- <pad> token included to make all captions of uniform length
- <unk> token added to represent rare words

Pre-processing (contd.)

3. Caption Pre-processing (contd.) - Example

Translated Captions: কুকুরটি উঠোন দিয়ে চলছে running

Noise Removed Captions: কুকুরটি উঠোন দিয়ে চলছে

Translated Captions: একজন মহিলা BESEIN রান্না করছে

Noise Removed Captions: একজন মহিলা রান্না করছে

Feature Extraction

1. **Image Feature Extraction**
   - 19-layer VGG, pre-trained on ImageNet dataset
   - 224 x 224 x 3 resized input image, 3 x 3 kernel with stride of 1 pixel
   - ReLU is used to introduce non-linearity
   - Pooling layers between convolutional layers, use max pooling over a (2 x 2) pixel window with a stride of 2
   - Among of the last three fully connected layers, output of the last fully connected layer before the classification layer was taken as features

Feature Extraction (contd.)

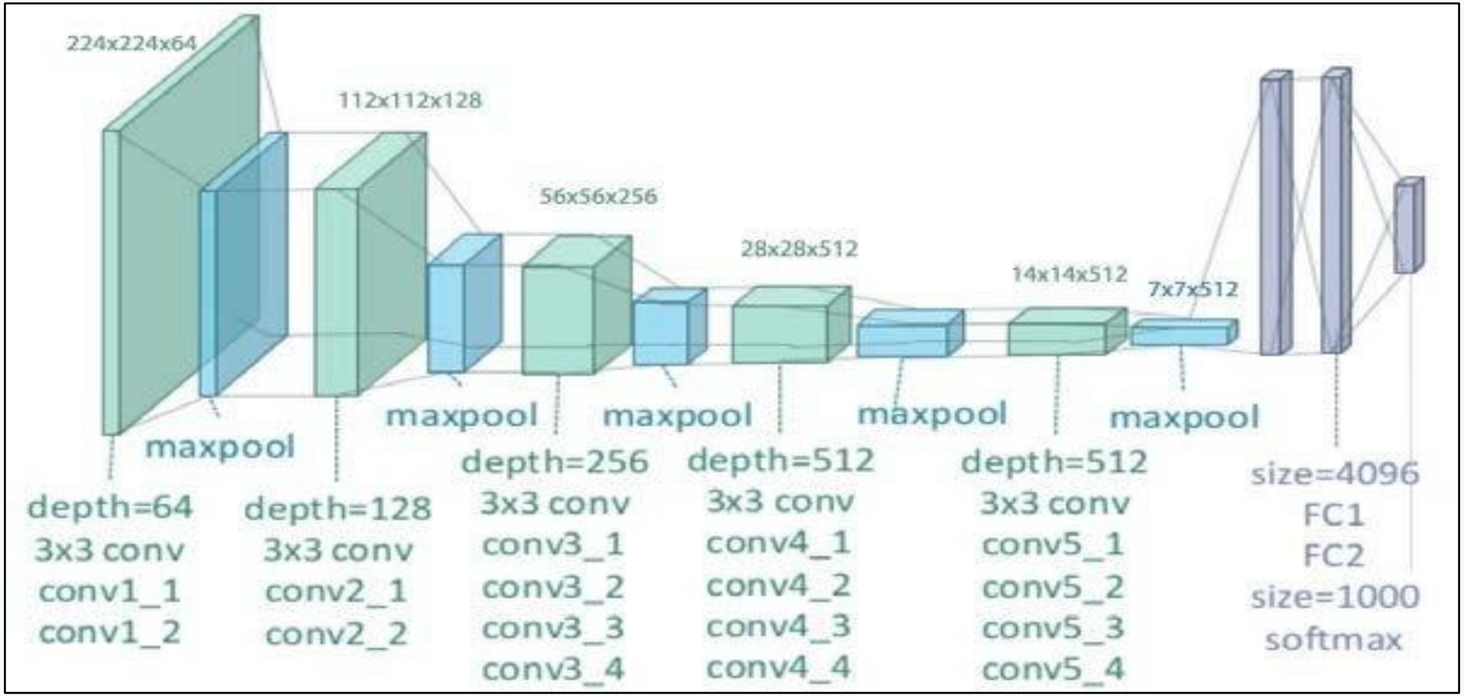


Figure: Main architecture of VGG-19

Figure taken from : Y. Zheng, C. Yang, and A. Merkulov, “Breast cancer screening using convolutional neural network and follow-up digital mammography,” p. 4, 05 2018.

Feature Extraction (contd.)

2. Video Feature Extraction

- ResNeXt-101, trained on Kinetics dataset, implemented to extract temporal motion features
- Activations of the last conv. layer extracted as the temporal feature representation for every 16 frames of a video.
- Extracted features were combined using max pooling.
- The last fully connected layer with softmax output is discarded.

Feature Extraction (contd.)

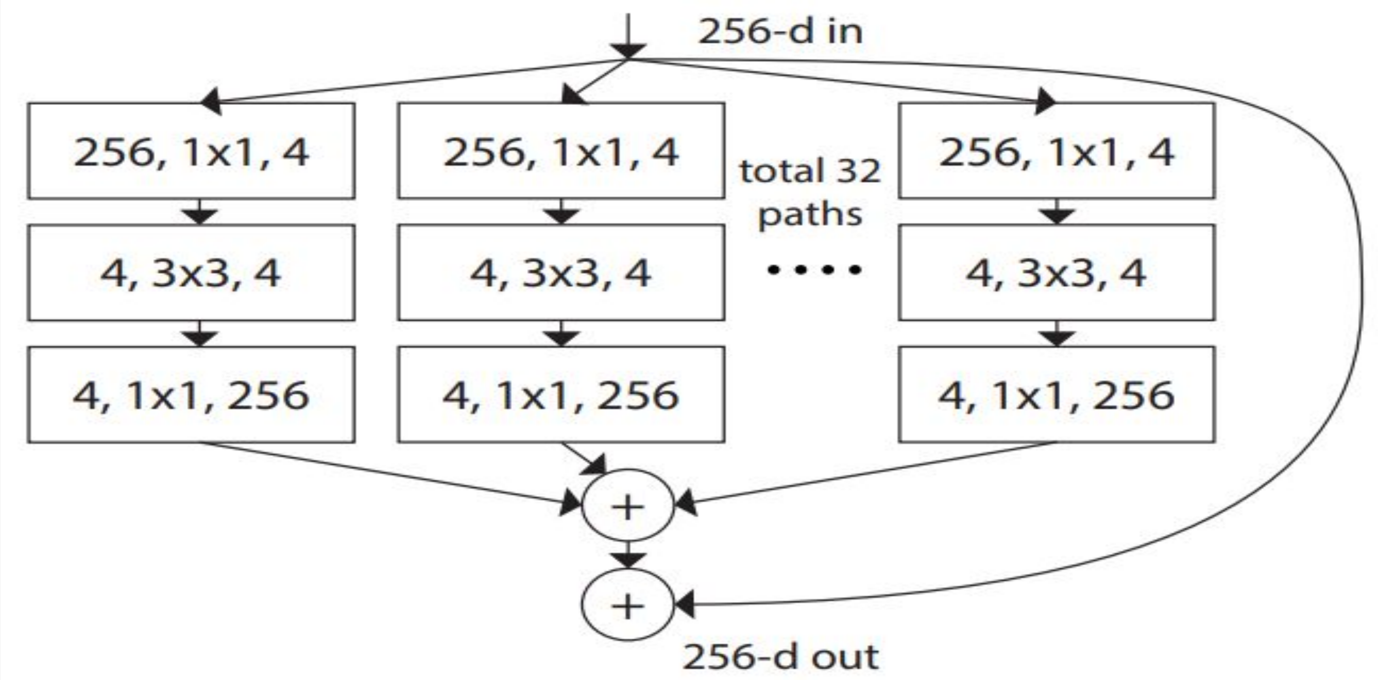


Figure: Single block of ResNeXt-101

Caption Embedding

3. Caption Pre-processing

- Used three different word embedding methods:
  - Word2Vec
  - FastText
  - GloVe
- Natural Language Processing (BNLP) toolkit is used
- Model trained with Bengali Wikipedia Dump Dataset
- Each word represented as a 300-dimension feature vector after embedding

Encoder-Decoder

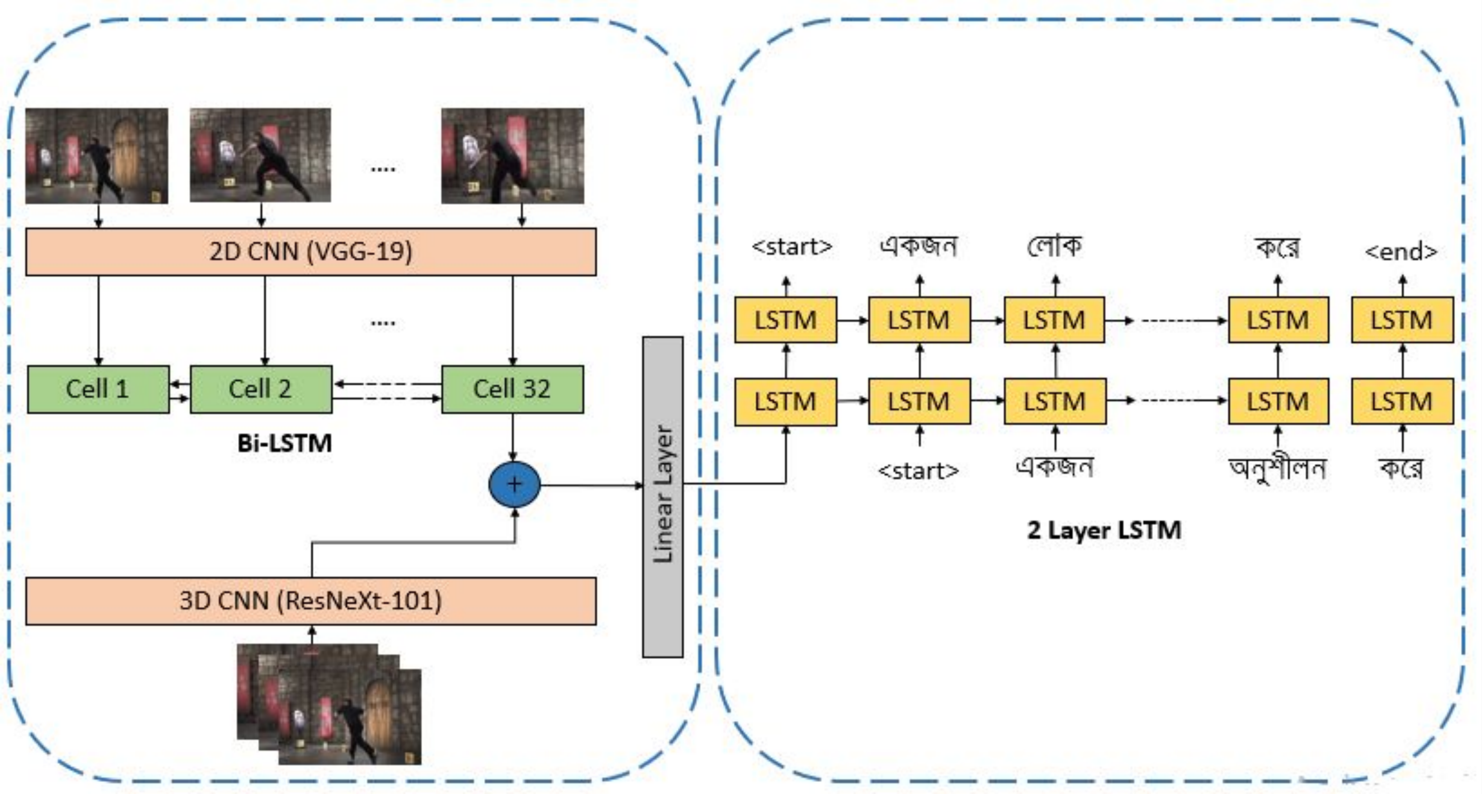


Figure: Proposed Encoder-Decoder Model

04 Results and Evaluation

Experimental Setup

- Used Python packages: Numpy, Pandas, OpenCV, H5py, bnltk etc.
- Image / video features are extracted from pre-trained models using torchvision.models PyTorch library
- Used ADAM (Adaptive Moment Estimation) optimization algorithm as optimizer and Cross-entropy loss function for loss calculation.
- Evaluated proposed model using three evaluation metric, BLEU, CIDEr, ROUGE

Hyper-parameters Setting

Uniform Hyper-parameters:

- Batch_size - 500
- Step_per_epoch - 99
- Epoch - 50
- Momentum - 0.0

Different Setups based on Non-uniform Parameters:

Setup	Encoder's Bi-LSTM Hidden Size	Encoder's LinearLayer Dropout	Decoder's LSTM Hidden Size	Decoder's LSTM Dropout	Word Embedding Dimension	Learning Rate
Setup-1	300	0.2	300	0.4	300	0.0005
Setup-2	500	0.2	500	0.4	500	0.005
Setup-3	1000	0.1	1000	0.5	1000	0.0002
Setup-4	1500	0.2	1500	0.4	1500	0.00005

Performance Comparison

Comparison Among Different Setups

1. Using FastText word embedding:

Setup	BLEU-3	BLEU-4	CIDEr	ROUGE
Setup - 1	**0.321**	**0.223**	0.276	0.74
Setup - 2	0.262	0.217	0.09	0.415
Setup - 3	0.308	0.221	**0.324**	**0.496**
Setup - 4	0.252	0.06	0.08	0.35

Performance Comparison (contd.)

Comparison Among Different Setups (contd.)

2. Using Word2Vec word embedding:

Setup	BLEU-3	BLEU-4	CIDEr	ROUGE
Setup - 1	0.286	0.220	0.314	0.502
Setup - 2	0.273	0.231	0.112	0.438
Setup - 3	**0.432**	**0.326**	**0.512**	**0.573**
Setup - 4	0.288	0.185	0.276	0.493

Performance Comparison (contd.)

Comparison Among Different Setups (contd.)

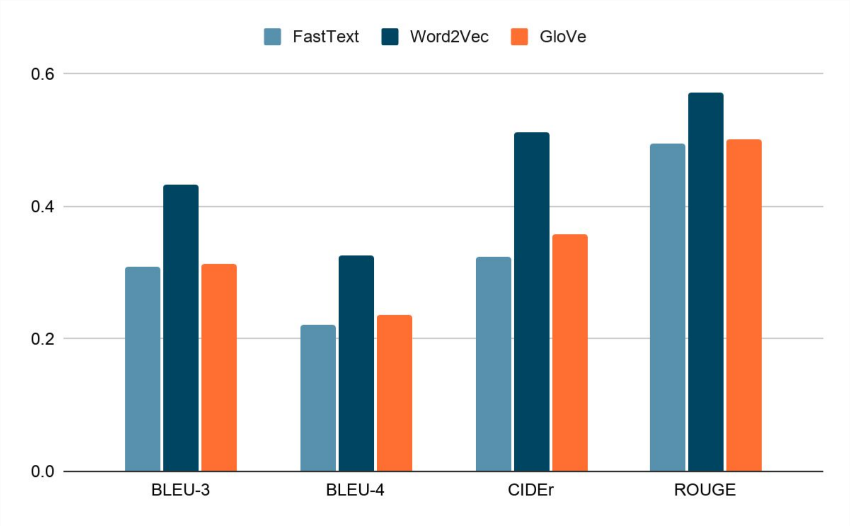
3. Using GloVe word embedding:

Setup	BLEU-3	BLEU-4	CIDEr	ROUGE
Setup - 1	**0.246**	**0.218**	0.253	0.458
Setup - 2	0.286	0.245	0.124	0.445
Setup - 3	0.314	0.237	**0.359**	**0.502**
Setup - 4	0.273	0.153	0.196	0.424

Performance Comparison (contd.)

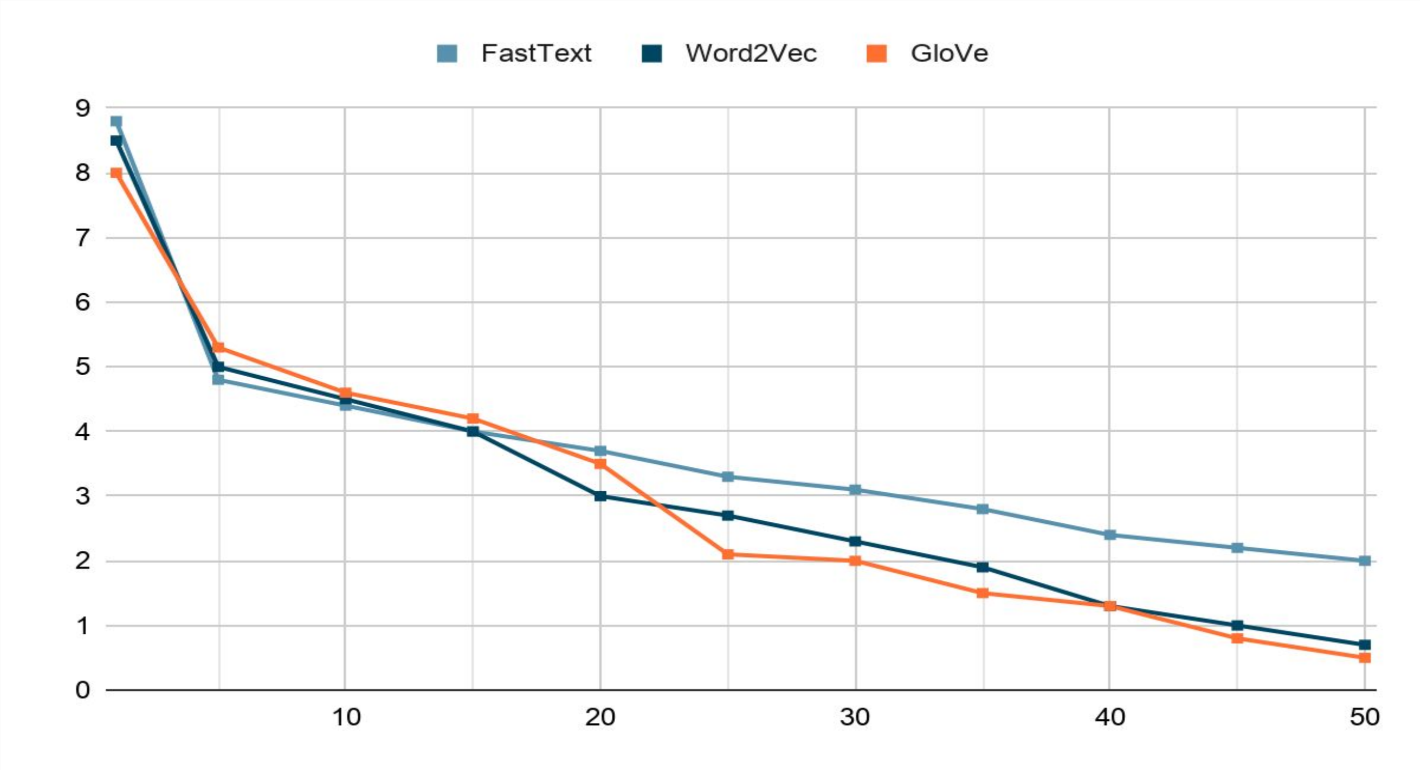
Performance of four metrics among three embedding model in Setup-3

Word Embedding Method	BLEU-3	BLEU-4	CIDEr	ROUGE
FastText	0.308	0.221	0.324	0.496
Word2Vec	**0.432**	**0.326**	**0.512**	**0.573**
GloVe	0.314	0.237	0.359	0.502



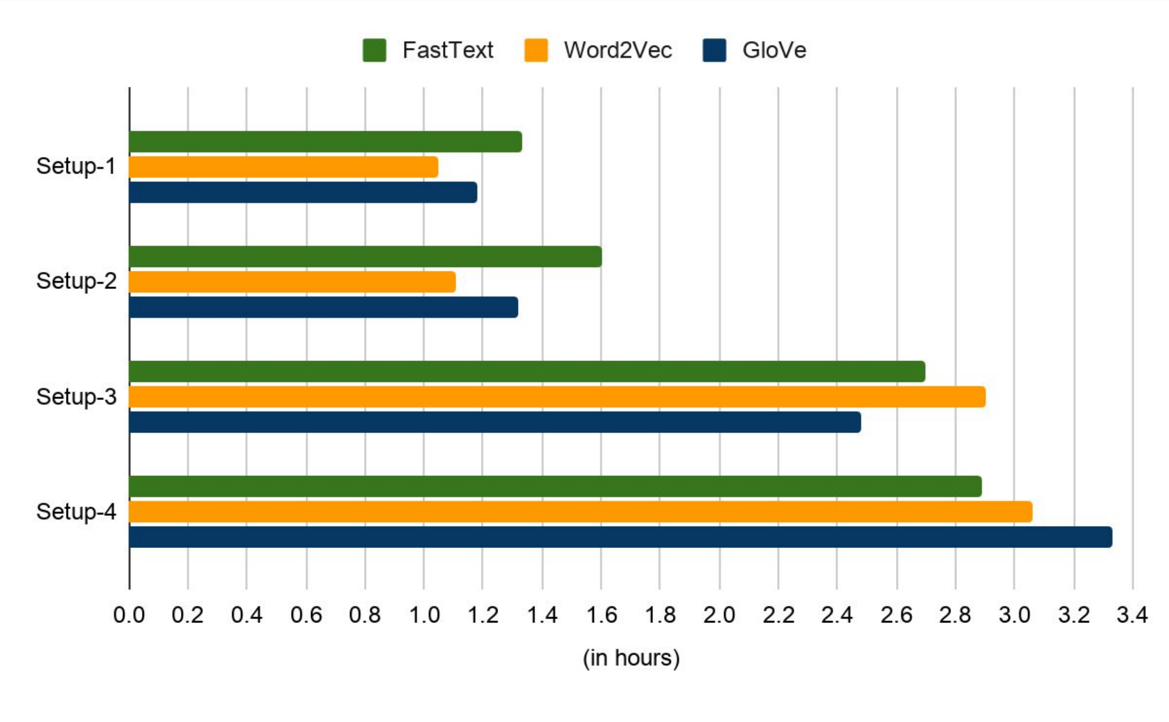
Performance Comparison (contd.)

Comparison of models' loss using different embedding model in Setup-3



Performance Comparison (contd.)

Comparison of times taken per epoch



Performance Comparison (contd.)

Comparison of times taken per epoch

Method	Dataset	BLEU-3	BLEU-4	CIDEr	ROUGE
Hybrid deep neural network[]	BNLIT (image)	32.4	22.8	-	-
CNN-RNN[2]	BanglaLekhaImageCaptions(image)	31.7	23.8	-	-
Par-inject and Merge architecture[3]	Flickr8k-BN (image)	33.0	22.0	46.0	54.0
Proposed Model	**MSVD (Video)**	**43.2**	**32.6**	**51.2**	**57.3**

[1] A. Jishan, K. R. Mahmud, A. K. Al Azad, S. Alam, and A. M. Khan, "Hybrid deep neural network for bangla automated image descriptor," International Journal of Advances in Intelligent Informatics, vol. 6, no. 2, pp. 109–122, 2020.
[2] A. H. Kamal, M. Jishan, N. Mansoor, et al., "Textmage: The automated bangla caption generator based on deep learning," arXiv preprint arXiv:2010.08066, 2020.
[3] T. Deb, M. Z. A. Ali, S. Bhowmik, A. Firoze, S. S. Ahmed, M. A. Tahmeed, N. Rahman, and R. M. Rahman, "Oboyob: A sequential-semantic bengali image captioning engine," Journal of Intelligent & Fuzzy Systems, no. Preprint, pp. 1–13, 2019.

05 Limitations and Future Works

Limitations

- No video dataset with Bengali captions
- Unable to handle complex visual information in videos (low accuracy)
- Short context vector leading to lack of knowledge of the complete context
- Huge computational expense

Future Works

- Use of attention mechanism to generate Bengali captions for better accuracy
- Preparing large video dataset with Bengali captions
- Resolving the problems with Bengali complex words and confusing meaning.
- Upgrade to description generation model

Thank You!

Feature Extraction (contd.)

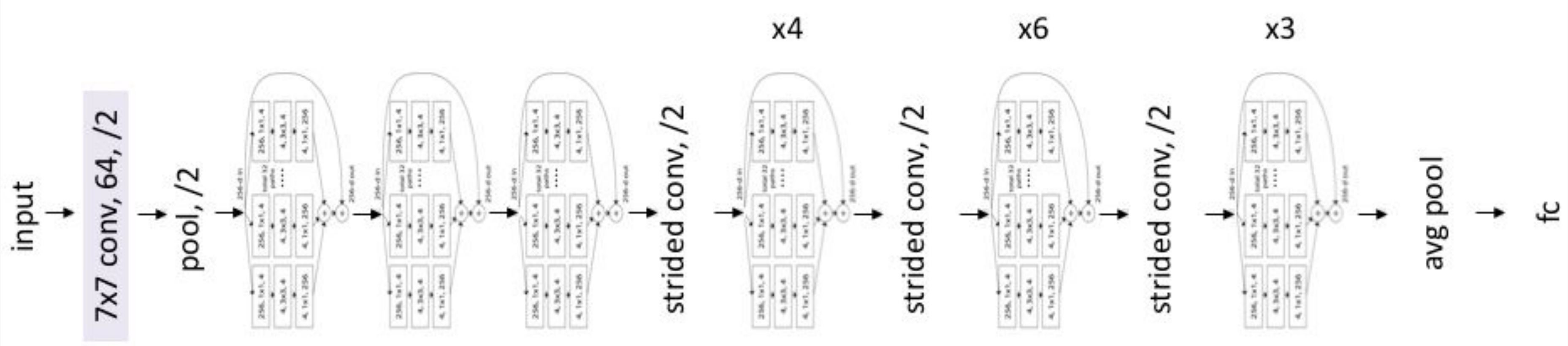


Figure: Main architecture of ResNeXt-101