


# Generating Bangla Image Captions with Deep Learning Techniques

Md. Anwar Hossain, Mirza AFM Rashidul Hasan, Sajeeb Kumar Ray, Naima Islam

# Table of Contents



Introduction
Materials
Methods
Results
Discussion

# Bangla Image Captioning



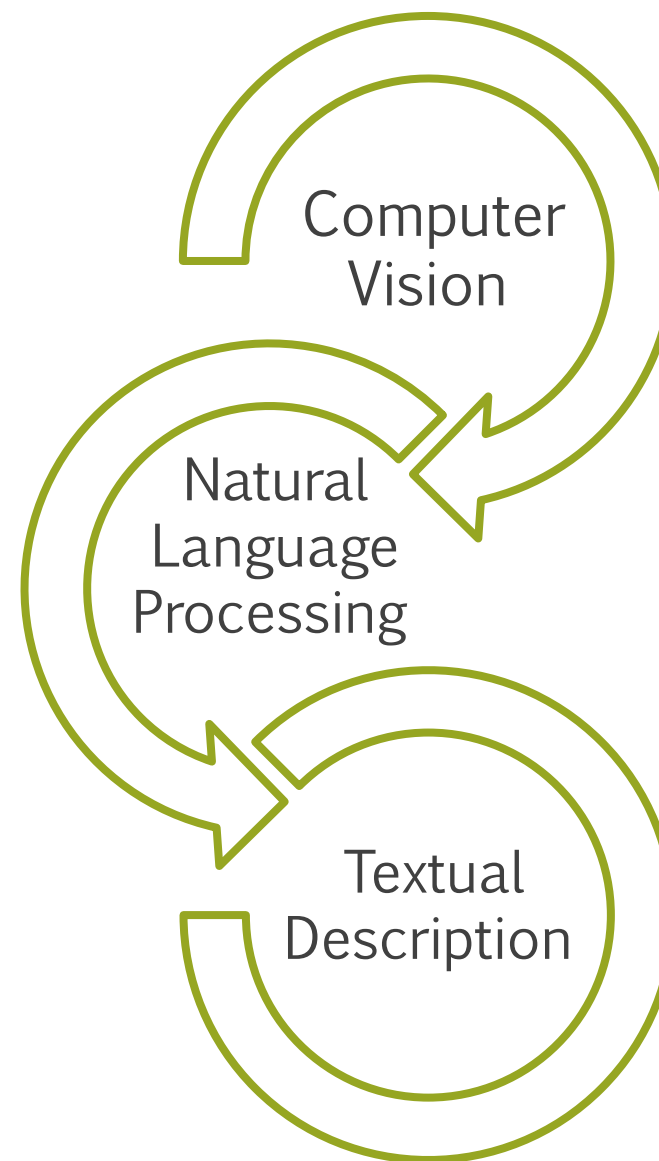
Input Image

## Automatic Text

একটি গোলাপী পোশাক  
পরা ছোট মেয়ে একটি  
কাঠের ঘরে যাচ্ছে

# Bangla Image Captioning

Bridging the gap  
between visual content  
and Bengali textual  
descriptions



# Our Contributions

- Developed a system for generating Bangla image captions.
- Employed deep learning techniques integrating computer vision and NLP.
- Used EfficientNetB4 and ResNet-50 for feature extraction.
- Introduced the BanglaView with the Flickr30k dataset for training and validation.



13691837.jpg



13880312.jpg



14072289.jpg



14107151.jpg



14133592.jpg



14220927.jpg



14252494.jpg



14264287.jpg



14377317.jpg



14526359.jpg



14580804.jpg



14698352.jpg



14698542.jpg



14748953.jpg



14799369.jpg



14868339.jpg



14887980.jpg



14919212.jpg



14955512.jpg



14989976.jpg

Images from  
A well known  
Flickr30k dataset  
Total image:  
31,783

332.jpg	0	একটি বড় হ্রদ যেখানে একটি একা হাঁস সাঁতার কাটছে এবং এর ধারে বেশ কিছু লোক রয়েছে
332.jpg	1	একটি ছোট ছেলে একটি সবুজ পার্ক দ্বারা ঘেরা জলে হাঁসের দিকে হাত নাড়ছে
332.jpg	2	দু'জন লোক একটি হ্রদের ধারে জল এবং শহরের আকাশরেখার মুখোমুখি
332.jpg	3	একটি বড় শহরে একটি শিশু এবং একজন মহিলা জলের ধারে
332.jpg	4	একটি লেকে একটি ছোট ছেলে হাঁস দেখছে
428.jpg	0	একটি দম্পতি এবং একটি শিশুকে একটি পুকুরের পাশে বসা পুরুষ দ্বারা আটকে রাখা হয়েছে একটি কাছাকাছি স্ট্রলার সহ
428.jpg	1	একজন পুরুষ এবং মহিলা জলের দেহের পাশে একটি শিশুর যত্ন নিচ্ছেন
428.jpg	2	এক দম্পতি তাদের নবজাতক শিশুকে নিয়ে একটি গাছের নিচে লেকের দিকে বসে আছেন
428.jpg	3	একটি শিশুর সাথে দম্পতি তাদের স্ট্রলারের পাশে বাইরে বসে আছে
428.jpg	4	একটি দম্পতি একটি শিশু এবং স্ট্রলার নিয়ে ঘাসের উপর বসে আছে
04.jpg	0	তিনজন লোক রাতে সাইকেল ভর্তি একটি আশ্রয় কেন্দ্রের কাছে একটি কালো গাড়ির কাছে দাঁড়িয়ে আছে
04.jpg	1	পার্ক করা গাড়ির পাশে একটি বিল্ডিংয়ের সামনে দাঁড়িয়ে কিছু লোক
04.jpg	2	আজ রাতে বাঁকা ছাদের নিচে অনেক বাইক পার্ক করা আছে
04.jpg	3	রাতে কালো ওয়াগনের চালককে জিজ্ঞাসাবাদ করছে পুলিশ
04.jpg	4	একটি কালো স্টেশন ওয়াগনের পাশে একটি গম্বুজের নীচে পুরুষ
728.jpg	0	ট্যাগ সহ একটি কালো ল্যাব জলে উল্লাস করছে
728.jpg	1	এটি একটি কালো কুকুর জলে ছিটকে পড়ছে
728.jpg	2	কালো কুকুরটি জলের মধ্যে দিয়ে চলে
728.jpg	3	একটি কালো কুকুর সার্ফের মধ্যে দৌড়াচ্ছে
728.jpg	4	একটি কুকুর জলে ছিটকে পড়ে

# Bangla captions from

Recently introduced  
**BanglaView** dataset

Total Caption:  
158,915



## Related Works

References	Dataset	Images	Captions
Khan et al. [1]	BanglaLekhalmageCaptions	9,154	18,308
Ami et al. [2]	Flickr8k, BanglaLekha, and Bornon	21,414	42,828
Bhuiyan et al. [3]	Bornon	4,100	20,500
Palash et al. [4]	BanglaLekhalmageCaptions	9,154	18,308
Faruk et al. [5]	BNATURE	8,000	40,000
Proposed	Flicker30k, BanglaView	31,783	1,58,915



# BanglaView: A Bangla Image Captioning Dataset



Vocabulary size:  
25,444 words



Captions per  
image: 5



Maximum length:  
67 words

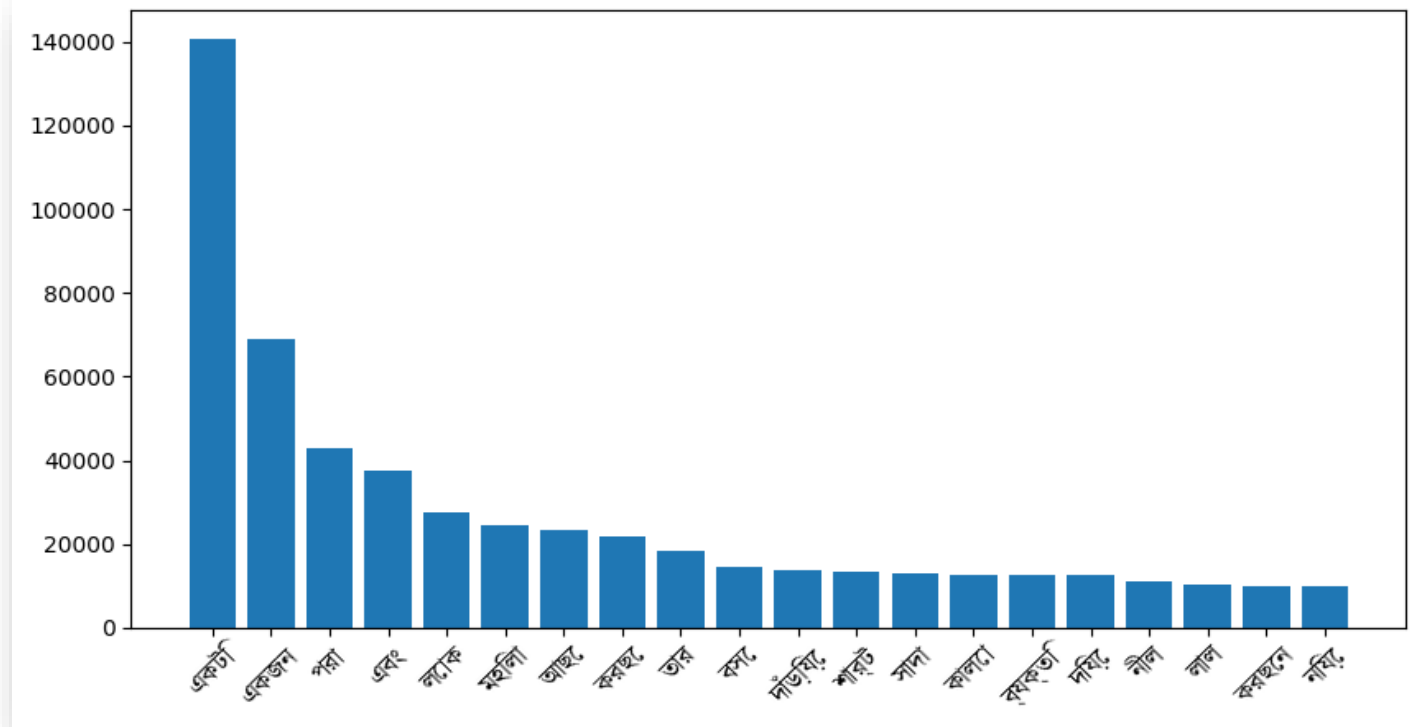
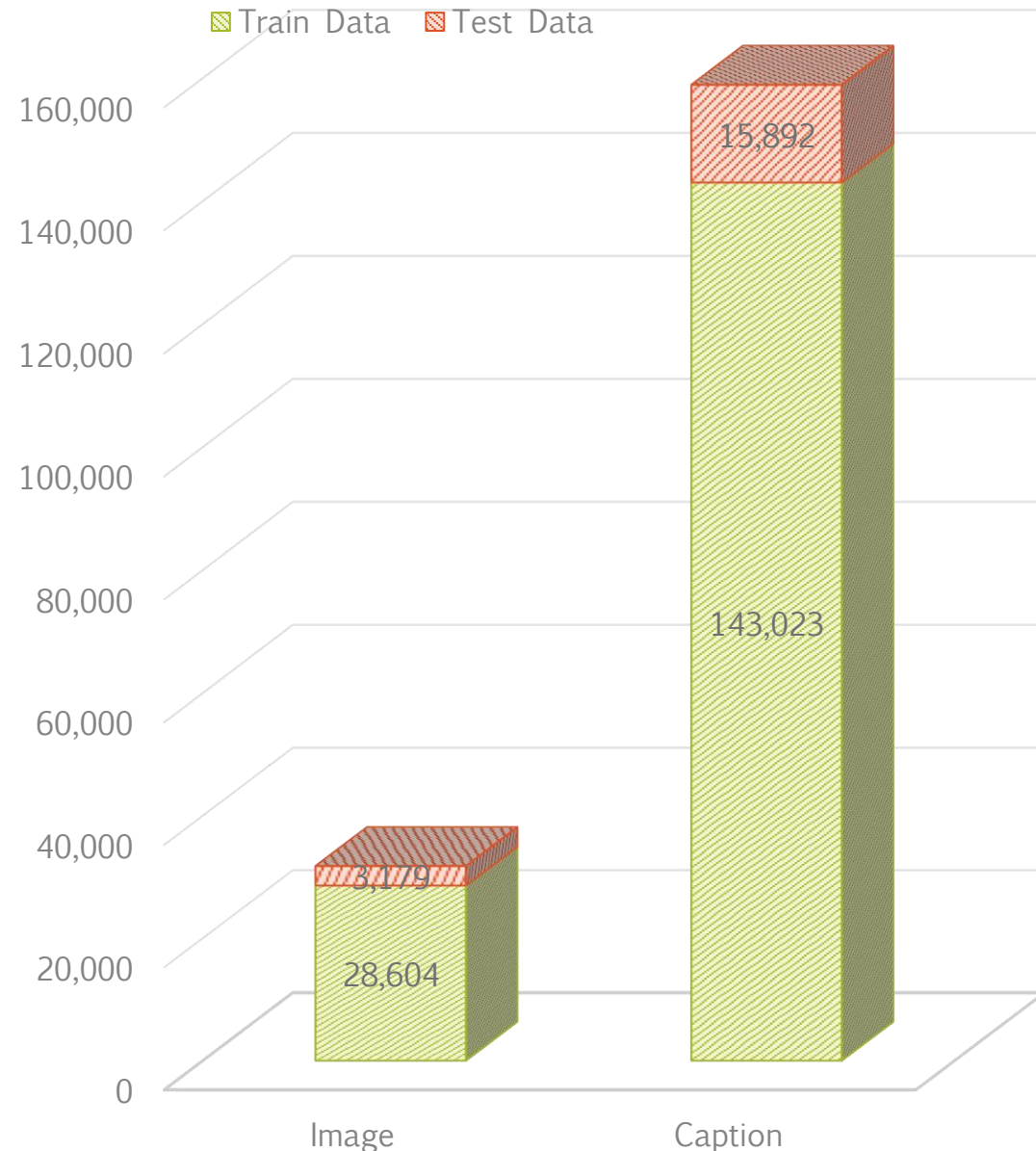


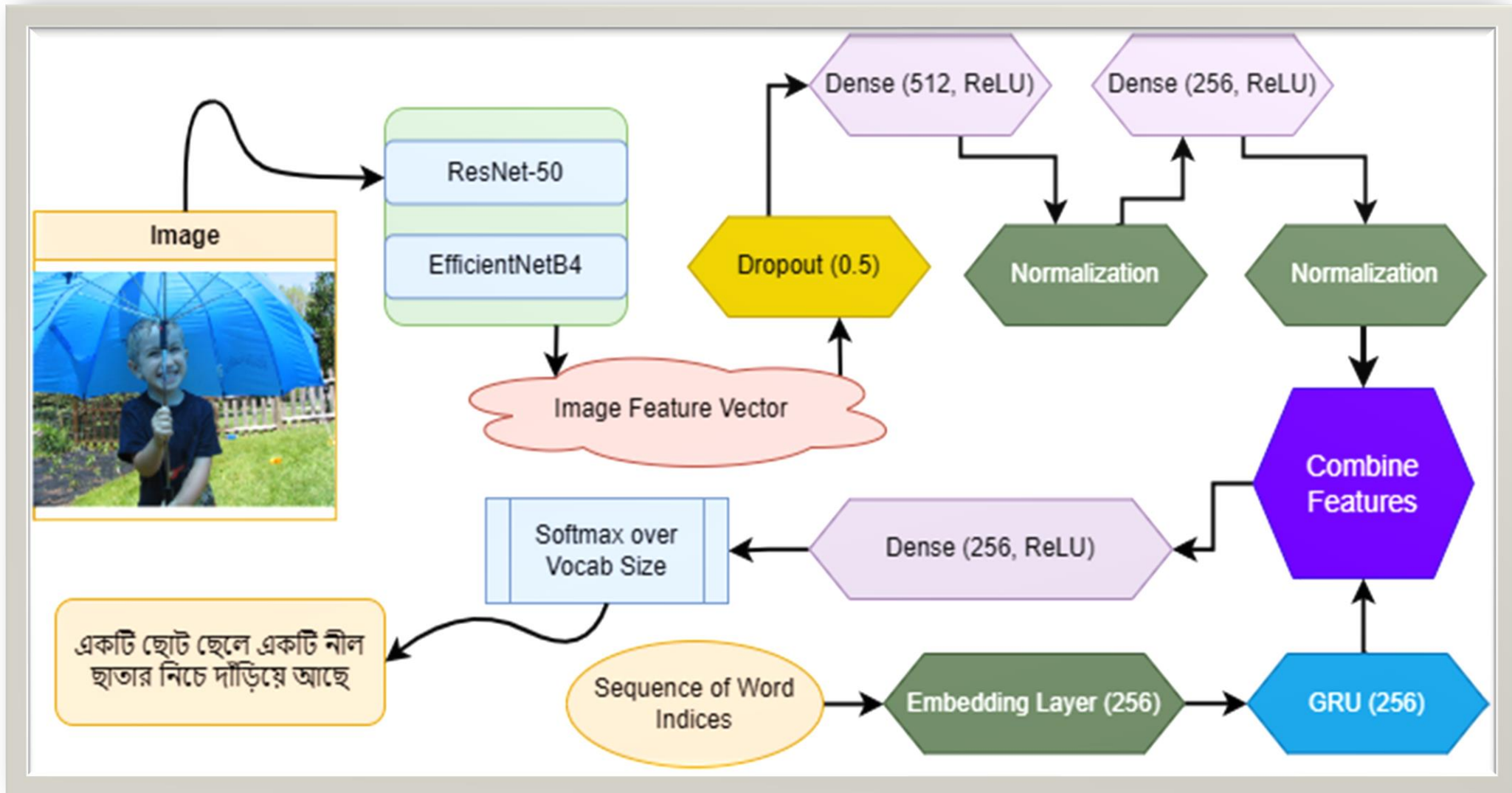
Figure: Most frequent words in BanglaView dataset

# Split Configuration

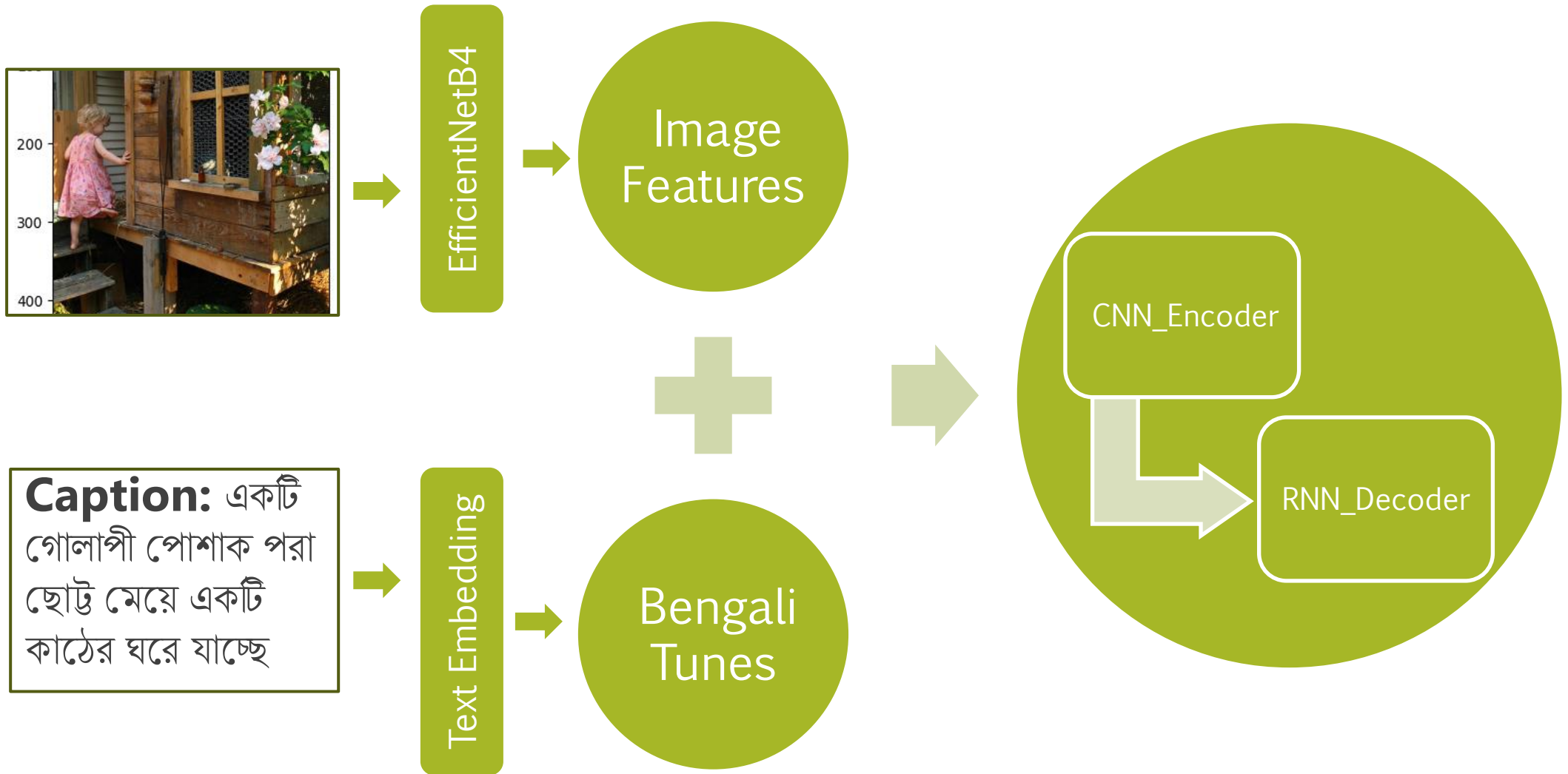
- Training Data (90%):  
28,604 images and  
143,023 captions
- Test Data (10%):  
3,179 images and  
15,892 captions



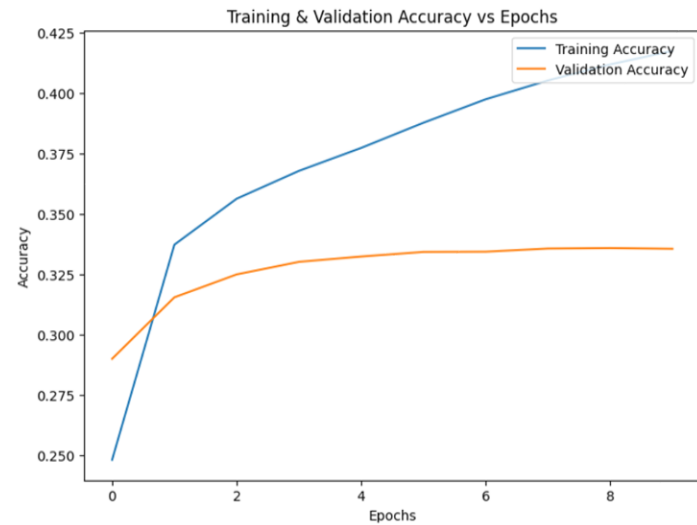
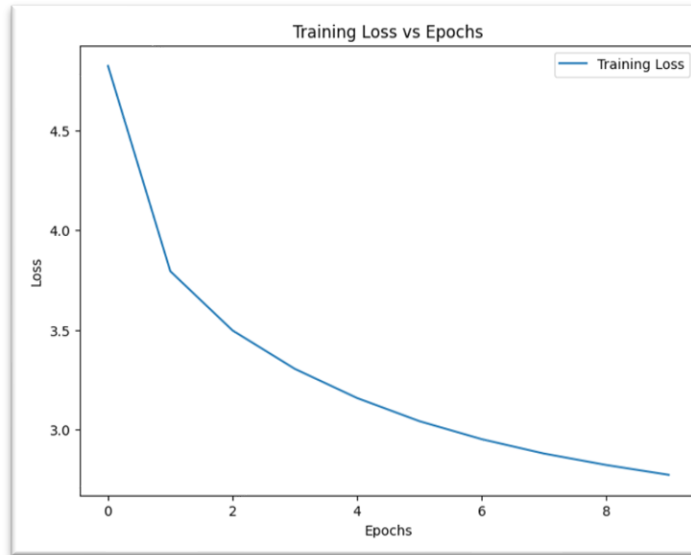
# System Architecture



# Training

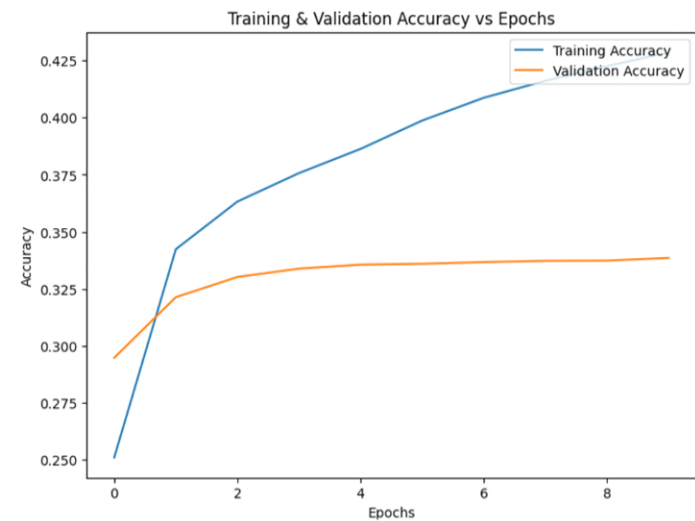
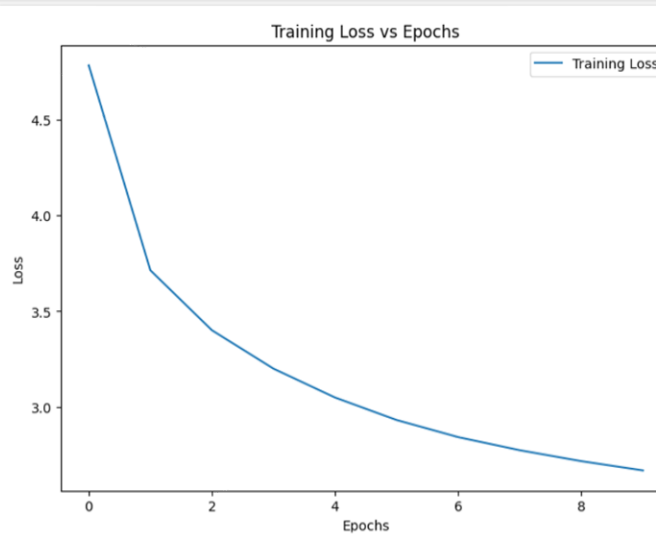


# Training Curves



ResNet50 training and validation graph

EfficientNetB4 training and validation graph



# Evaluation (After 10 Epochs of Training)

## Test Scores

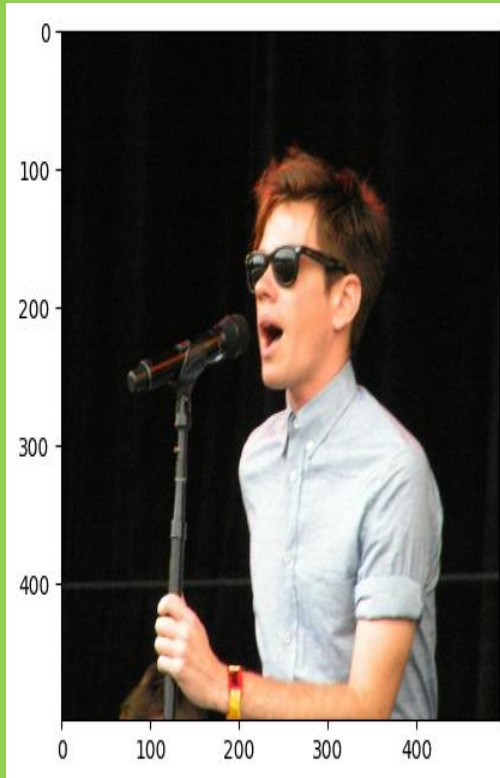
Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
EfficientNetB4	0.5396	0.3442	0.2044	0.1163	0.9896	0.9621	0.9893
ResNet-50	0.5011	0.3101	0.1763	0.0944	0.9861	0.9517	0.9861

## BLUE Scores and its Meaning

Score Range	Quality Description	Score Range	Quality Description
0.00 - 0.10	Very Poor	0.50 - 0.60	Excellent
0.10 - 0.20	Poor	0.60 - 0.70	Outstanding
0.20 - 0.30	Fair	0.70 - 0.80	Near Perfect
0.30 - 0.40	Good	0.80 - 0.90	Exceptional
0.40 - 0.50	Very Good	0.90 - 1.00	Perfect

# Evaluation

Flicker30k Image



## BanglaView Captions

C1: একটি বোতামযুক্ত শার্ট এবং ভিনটেজ-স্টাইলের সানগ্লাস পরে মঞ্চে একজন তরুণ অভিনয়শিল্পী তার গান গাইছেন (A young performer wearing a buttoned shirt and vintage-style sunglasses is singing his song on stage)

C2: সানগ্লাস এবং হালকা নীল শার্ট-হাতা শার্ট পরা একজন শ্যামাঙ্গিনী ব্যক্তি একটি মাইক্রোফোনে গান গাচ্ছেন

C3: হালকা নীল শার্ট পরা একজন যুবক মাইক্রোফোনে কথা বলছে বা গান করছে

C4: সানগ্লাস পরা একজন সাদা লোক মাইক্রোফোন নিয়ে গান গাইছে

C5: একজন তীক্ষ্ণ পোশাক পরা লোক মাইক্রোফোনে গান গাইছে

## Model Generated

EfficientNetB4: একজন লোক মাইক্রোফোনে গান গাইছে (A man is singing into a microphone)

ResNet50: একটি কালো এবং সাদা মাইম এবং একটি সাদা শার্ট পরা একজন ব্যক্তি একটি মাইক্রোফোনে গান গাইছেন



# Comparative Study

References	Model	BLUE-1
Khan et al. [1]	CNN-Merge	0.65
Ami et al. [2]	Visual-Attention	0.59
Bhuiyan et al. [3]	ResNet50+Attention+BIGRU	0.85
Palash et al. [4]	ResNet101	0.69
Faruk et al. [5]	CNN + RNN	0.43
Proposed	CNN + RNN	0.54

# Applications

1



Accessibility for  
visually  
impaired  
individuals

2



Photo search  
enhancement

3



Robot  
Interaction

# References

- [1] Faiyaz Khan, M., Sadiq-Ur-Rahman, S., & Saiful Islam, M. "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020
- [2] Ami, A. S., Humaira, M., Jim, M. A. R. K., Paul, S., & Shah, F. M. "Bengali image captioning with visual attention," 2020 23<sup>rd</sup> International Conference on Computer and Information Technology (ICCIT)
- [3] Ahatesham Bhuiyan, Eftekhar Hossain, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Enhancing image caption generation through context-aware attention mechanism, Heliyon, Volume 10, Issue 17, 2024, e36272, ISSN 2405-8440
- [4] Palash, M.A.H., Nasim, M.A.A., Saha, S., Afrin, F., Mallik, R., Samiappan, S. (2022). Bangla Image Caption Generation Through CNN-Transformer Based Encoder-Decoder Network. In: Hossain, S., Hossain, M.S., Kaiser, M.S., Majumder, S.P., Ray, K. (eds) Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021 . Lecture Notes in Networks and Systems, vol 437. Springer, Singapore.
- [5] A. M. Faruk, H. A. Faraby, M. M. Azad, M. R. Fedous and M. K. Morol, "Image to Bengali Caption Generation Using Deep CNN and Bidirectional Gated Recurrent Unit," 2020 23<sup>rd</sup> International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-6



**THANK YOU!**