Project

Deep Learning [CSE477s]

Contents

1) Team Members 4
2) Github Link 6
2.1) Github Link 7
3) Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks 8
3.1) Overview & Problem Statement 9
3.2) Objectives & Achievements 9
3.2.1) Outcome 10
3.3) Methodology 10
3.3.1) Model: Show, Attend and Tell 10
3.3.2) Loss Functions Compared 10
3.3.3) Optimizers Compared 11
3.4) Other Applicable Methods 12
3.5) Preferred Alternatives & Rationale 13
3.6) Dataset & Generalizability 13
3.6.1) Dataset Used 13
3.6.2) Generalization Challenges 14
3.6.3) Suggested Improvements 14

Contents (ii)

3.7) Results & Comparative Analysis 15
3.7.1) Paper Results Summary 15
3.7.2) Comparison with Other Methods 16
3.8) Critique & Suggestions 17
3.8.1) Criticisms 17
3.8.2) Suggested Enhancements 18
3.9) Personal Implementation 18
3.9.1) Implementation Summary 18
3.9.2) Result Comparison 19
3.9.3) Reasons for Gaps 21
3.10) Screenshots of Sample Outputs 22
3.11) Training Loss Plot 23
3.12) Sample Captions 24
4) References 27
4.1) References 28

1) Team Members

Contents (iv)

Name	ID
Mostafa Hamada Hassan Ahmed	2100587
Zeyad Magdy Roushdy	2100393
Mina Nasser Shehata	2100370
Mina Antony Samy Fahmy	2101022

2) Github Link

Github Link

2.1) Github Link

You can find our Github repository on the following link:

- https://github.com/7ksha/Image_Captioning-

3) Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks

Objectives & Achievements

3.1) Overview & Problem Statement

Image captioning aims to generate natural language descriptions from images. This paper explores:

- The Show, Attend and Tell model (CNN + LSTM + Attention)
- Effects of different loss functions (cross-entropy, MSE, KL divergence, NLL)
- Comparisons of optimizers (Adam, SGD, RMSProp, Adadelta)

This work contributes to AI systems capable of interpreting and describing visual data:

> *Building AI systems that can understand images and describe them in human language*

3.2) Objectives & Achievements

- Evaluate loss functions and optimizers for image captioning tasks
- Study effects of hyperparameter tuning
- Replicate the Show, Attend and Tell model
- Offer practical insights for choosing configurations

3.2.1) Outcome

✅ Achieved: Comparison of functions/optimizers, identified best combination (Adam + cross-entropy).

❌ Limitations: Evaluated only one model; no transformer-based models compared.

3.3) Methodology

3.3.1) Model: Show, Attend and Tell

- **Encoder**: CNN extracts image features
- **Decoder**: LSTM generates captions with soft attention
- **Training**: Used teacher forcing

3.3.2) Loss Functions Compared

- Cross-Entropy
- KL Divergence
- Mean Squared Error
- Negative Log-Likelihood

Methodology (ii)

3.3.3) Optimizers Compared

- Adam
- RMSProp
- SGD
- Adadelta

Other Applicable Methods

3.4) Other Applicable Methods

METHOD	APPLICATION
Vision Transformers (ViT)	Patch-based input encoding + transformer decoder
DeiT	Efficient variant of ViT
BLIP / BLIP-2	Multimodal vision-language transformers
SCST	Reinforcement learning-based caption refinement
UpDn	Uses object detection features
CLIPCap	Zero-shot captioning via CLIP
Hybrid Models	CNN-Transformer combination
Contrastive Learning	Cross-modal representation learning

Dataset & Generalizability

3.5) Preferred Alternatives & Rationale

Transformer-based models (ViT, BLIP) are preferred for:

- Better scalability
- Stronger long-range dependency modeling
- Higher accuracy and generalization

LSTM-based models are simpler and good for prototyping but limited in complexity and scope.

3.6) Dataset & Generalizability

3.6.1) Dataset Used

- MS COCO 2014: 120k images with 5 captions each

Dataset & Generalizability (ii)

3.6.2) Generalization Challenges

DOMAIN	GENERALIZATION RISK
Medical/Satellite Images	❌ Poor generalization
Low-resource Languages	❌ Poor performance
Smaller Datasets	❌ Less robust
Clinical Applications	❌ Not directly applicable

3.6.3) Suggested Improvements

- Test on datasets like Flickr8k/Flickr30k
- Apply domain adaptation or transfer learning
- Add metadata features (age, context, lighting)

3.7) Results & Comparative Analysis

3.7.1) Paper Results Summary

MODEL	RESULT
Best Optimizer	Adam
Best Loss Function	Cross-Entropy
Evaluation Metrics	Top-5 Accuracy, BLEU-4
Model	Show, Attend and Tell

~27–29% BLEU-4 estimated based on architecture.

Results & Comparative Analysis (ii)

3.7.2) Comparison with Other Methods

MODEL	DATASET	BLEU-4	NOTES
Show, Attend and Tell	MS COCO	27%	Baseline
UpDn	MS COCO	36%	Uses object detection
SCST	MS COCO	39%	Reinforcement learning
BLIP	MS COCO	45%	Transformer-based
CLIPCap	MS COCO	42%	Zero-shot capable
User Implementation	Flickr8k	9.71%	Educational demo

Critique & Suggestions

3.8) Critique & Suggestions

3.8.1) Criticisms

- Narrow model choice (only Show, Attend and Tell)
- No transformer or hybrid models
- Only used MS COCO
- Lacks newer metrics (CIDEr, SPICE)
- No qualitative examples (e.g., attention maps)

Personal Implementation

3.8.2) Suggested Enhancements

IMPROVEMENT	BENEFIT
Use ViT/DeiT	Higher accuracy & generalization
Add examples & maps	Interpretability
Add beam search	Better output quality
Report CIDEr/METEOR	Improved evaluation
Use more/larger datasets	Better robustness

3.9) Personal Implementation

3.9.1) Implementation Summary

- **Encoder**: EfficientNetB7
- **Decoder**: LSTM with attention
- **Dataset**: Flickr8k
- **Training**: Cross-entropy + Adam + teacher forcing

Personal Implementation (ii)

3.9.2) Result Comparison

Aspect	Paper (Show, Attend and Tell + ResNet/VGG)	Your Implementation
Best Optimizer	Adam	✅ Used Adam
Best Loss Function	Cross-Entropy	✅ Used Cross-Entropy
BLEU-4 Score	19–34%	❌ Achieved 9.71%
METEOR	Not explicitly reported	✅ Measured but low
CIDEr	Not reported	✅ Measured
Caption Quality	High semantic accuracy	Fair, some redundancy
Encoder Used	VGG-16 / ResNet	EfficientNetB7
Attention Visualization	Yes	❌ Not yet implemented
Beam Search	Yes	❌ Currently using greedy search

Personal Implementation (iii)

Aspect	Paper (Show, Attend and Tell + ResNet/VGG)	Your Implementation
Multi-GPU Support	Implied	❌ Single GPU used
Training Time	Longer (full COCO)	Shorter (subset used)

Personal Implementation (iv)

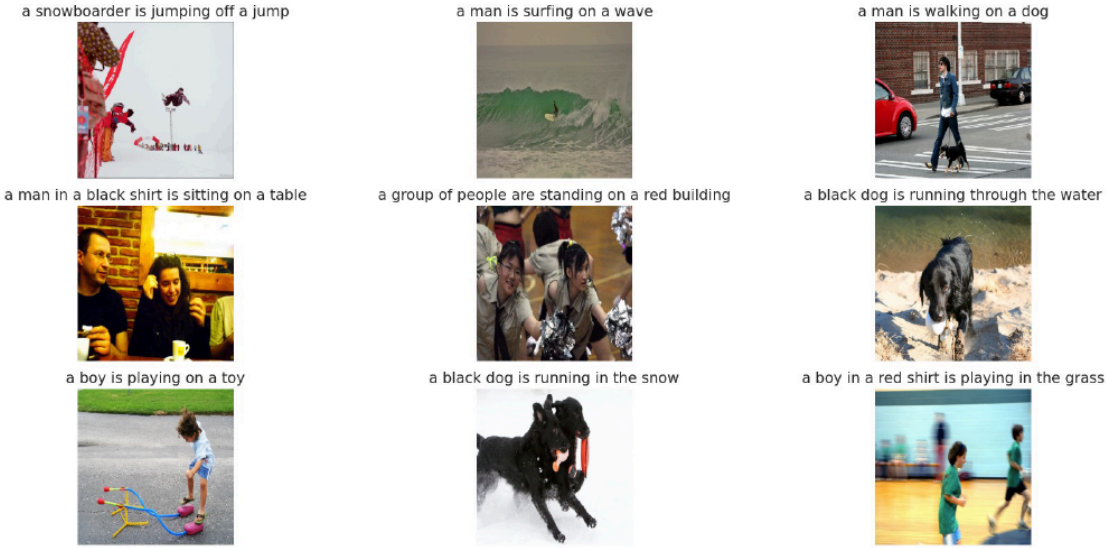
3.9.3) Reasons for Gaps

DIFFERENCE	REASON
Smaller Dataset	Flickr8k vs MS COCO
Shorter Training	Fewer epochs
Greedy Decoding	No beam search
No Attention Visualization	Omitted
Simplified Hyperparams	Limited resources

Screenshots of Sample Outputs

3.10) Screenshots of Sample Outputs

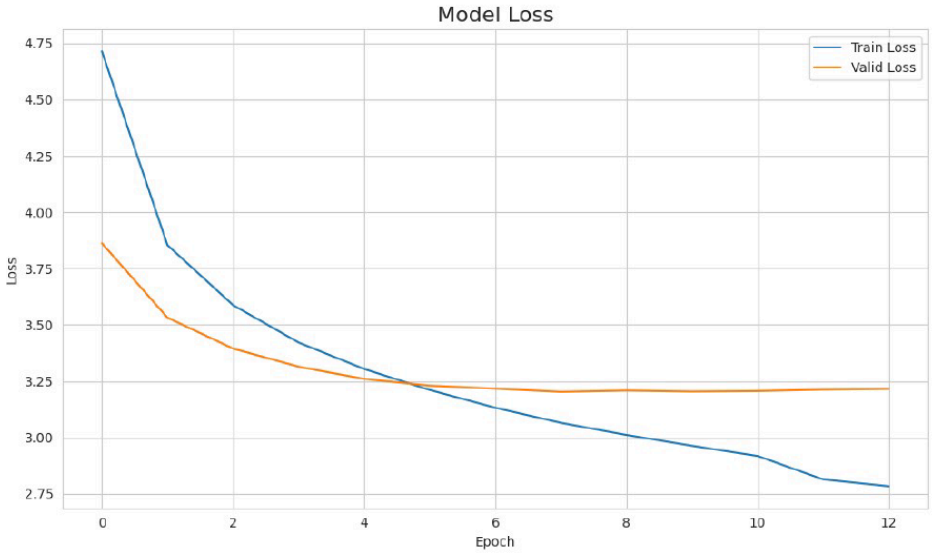
Example format:



Training Loss Plot

3.11) Training Loss Plot

Plot showing training vs validation loss:



Sample Captions

3.12) Sample Captions

Image	Ground Truth Caption	Generated Caption
	A child climbing stairs	A girl climbing up the stairs

Sample Captions (ii)

Image	Ground Truth Caption	Generated Caption
	A black dog and a spotted dog are fighting	Two dogs are fighting on the road

Sample Captions (iii)

Image	Ground Truth Caption	Generated Caption
	A man riding skateboard	A man is riding a skateboard on pavement

4) References

References

4.1) References

[1] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[2] "[1502.03032)] Article identifier not recognized." [Online]. Available: https://arxiv.org/abs/1502.03032)

[3] "Self-critical Sequence Training for Image Captioning." [Online]. Available: http://arxiv.org/abs/1612.00563

[4] "Baby talk: Understanding and generating simple image descriptions," in *CVPR 2011*, Jun. 2011, pp. 1601–1608. doi: 10.1109/CVPR.2011.5995466.

[5] "Corpus-Guided Sentence Generation of Natural Images," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 444–454. [Online]. Available: https://aclanthology.org/D11-1041/

References (ii)

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." [Online]. Available: http://arxiv.org/abs/1905.11946

[7] "NLTK :: Natural Language Toolkit." [Online]. Available: https://www.nltk.org/

[8] "PyCOCOEvalCap – Automatic Caption Evaluation - Google Search." [Online]. Available: https://www.google.com/search?client=firefox-b-d&q=PyCOCOEvalCap+%E2%80%93+Automatic+Caption+Evaluation

[9] "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." [Online]. Available: http://arxiv.org/abs/1502.03044