Computer and Systems Engineering [CSE]

Deep Learning [CSE477s]

Project Questions

Team Members

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

Contents

1) What is the problem statement of the paper?

The paper titled "Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks" addresses the challenge of automatically generating natural language descriptions from images , otherwise known as image captioning . The specific goals of the study are:

- To evaluate the Show, Attend and Tell architecture
- To analyze the impact of different loss functions (e.g., cross-entropy, KL divergence, mean-squared error, negative log-likelihood)
- To compare various optimizers (Adam, RMSProp, SGD, Adadelta) in training attention-based image captioning models

The broader research context aims to build vision-language AI systems that can understand visual data and describe it using human-readable text — an essential task in fields like assistive technology, robotics, and medical imaging. This aligns with the general vision-language task:

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

2) What are the objectives of the paper and do you think the authors managed to achieve these goals? Explain

2.1) Objectives

1. Evaluate different loss functions and optimizers on the Show, Attend and Tell model for image captioning: Compare performance of different combinations (MSE, NLL, KL vs Adam, SGD, RMSProp, Adadelta).
2. Study the effect of hyperparameter tuning: Analyze how learning rate, batch size, and other parameters affect BLEU-4 and Top-5 Accuracy.
3. Replicate Show, Attend and Tell Architecture: Use CNN + LSTM + Attention mechanism.
4. Provide Practical Insights: Help practitioners choose optimal configurations for image captioning tasks.

2.2) Did they achieve these goals?

Yes, the authors largely achieved their stated objectives:

- They conducted experiments using multiple loss functions and optimizers on the MS COCO dataset
- They measured performance using standard metrics such as Top-5 Accuracy and BLEU-4
- They reported results showing that certain combinations (especially cross-entropy + Adam) yield better performance
- They concluded that Adam optimizer with cross-entropy loss gave the best overall results

However, the study was somewhat limited in scope:

- Only one model architecture (CNN + LSTM + Attention) was tested
- No comparison with more recent transformer-based methods (e.g., Vision Transformers or DeiT)

So while the authors met their main goals, they didn't explore cutting-edge architectures.

3) What is the DL method used in this paper?

The paper implements and evaluates the Show, Attend and Tell architecture, which is a deep learning-based encoder-decoder framework with soft visual attention.

3.1) Model Breakdown

Encoder:

- Based on a Convolutional Neural Network (likely ResNet or VGG)
- Extracts spatial feature vectors from input images
- Output shape: (batch_size, num_regions, feature_dim) (e.g., 7×7 grid with 2048 features per region)

3.2) Decoder

- Uses an LSTM network for sequence generation
- Applies soft attention over the encoder's output at each decoding step
- Predicts next word using softmax over vocabulary

3.3) Training Strategy

- Used teacher forcing: Feeding ground-truth words during training
- Four loss functions were compared:
  - Cross-Entropy
  - Kullback-Leibler Divergence
  - Mean Squared Error
  - Negative Log-Likelihood

3.4) Optimizers Compared

- Adam
- RMSProp
- SGD
- Adadelta

The authors concluded that Adam with cross-entropy loss performed best, although they acknowledged that newer transformer-based approaches would likely outperform their baseline model.

4) What are the other state-of-the-art methods that can be applied to the same problem?

Several advanced deep learning methods could also be applied to image captioning:

METHOD	APPLICATION
Vision Transformers (ViT)	Converts images into patches, encodes them, and uses a transformer decoder for captioning
DeiT (Data-efficient Image Transformer)	Lightweight variant of ViT, designed for small datasets
BLIP / BLIP-2	Multimodal models trained on large-scale image-text pairs; excellent zero-shot capability
Self-Critical Sequence Training (SCST)	Reinforcement Learning approach that improves CIDEr score by learning from generated captions
UpDn (Anderson et al.)	Uses object detection (Faster R-CNN) to extract ROI features first
CLIPCap	Leverages CLIP embeddings and fine-tunes the language side only
Hybrid Models	Combine CNNs for local features and Transformers for global context modeling
Contrastive Learning Methods (CrossCLR, CoCLR)	Learn joint representations between modalities (useful for multi-modal video/ image captioning)

These methods offer advantages such as better accuracy, efficiency, or interpretability depending on the specific use case.

5) Would you apply any of the other methods other than the DL method used in this paper? Explain your answer

Yes, I would consider applying other methods , especially those based on transformer architectures or multi-modal learning. Reasons:

1. Vision Transformers (ViT / DeiT):
   - Better at capturing long-range dependencies
   - Can be pre-trained on large datasets like Conceptual Captions or LAION
   - Often outperform CNN+LSTM baselines
2. BLIP / BLIP-2:
   - Combines vision and language transformers
   - Capable of both captioning and visual question answering (VQA)
3. Self-Critical Sequence Training (SCST):
   - Significantly improves CIDEr and BLEU scores
   - Learns from its own predictions instead of just ground truth
4. LSTM with Attention (what I implemented):
   - Simpler and faster to train
   - Good for prototyping
   - Less scalable than transformer-based models

Conclusion: While the method in the paper provides a strong baseline , I would prefer to explore transformer-based models (like ViT or DeiT) for higher performance and scalability in future work.

6) What datasets have been used in this paper? Do you think the result is generalizable for any datasets?

6.1) Dataset Used

The paper utilized the MS COCO dataset (2014 version)

- Contains over 120,000 annotated images
- Each image has five captions
- Standard benchmark for image captioning tasks

6.2) Generalizability

Although MS COCO is widely accepted and standardized:

DOMAIN	GENERALIZATION RISK
Specialized domains	❌ Medical imaging, satellite imagery, etc.
Low-resource languages	❌ Non-English speakers or underrepresented languages
Smaller or less varied datasets	❌ Flickr8k, Pascal VOC
Real-world clinical applications	❌ Not directly applicable without domain adaptation

6.3) Improvements for Generalization

- Test on Flickr8k/Flickr30k and compare performance
- Apply transfer learning or domain adaptation techniques
- Incorporate metadata (age, gender, lighting conditions) if available

7) Discuss the results presented in the paper. Compare the results with other state-of-the-art methods used to solve this problem.

7.1) Results from the Paper

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

No exact BLEU-4 score was reported, but based on similar studies and evaluations, we can estimate that the CNN+Attention setup reached approximately ~27–29% BLEU-4 score .

7.2) Comparison with Other State-of-the-Art Methods

MODEL	DATASET	BLEU-4	NOTES
Show, Attend and Tell (CNN + Attention)	MS COCO	~27%	Baseline model
UpDn (Anderson et al.)	MS COCO	~36%	Uses object detection features
SCST (Rennie et al.)	MS COCO	~39%	Uses reinforcement learning
BLIP (Li et al.)	MS COCO	~45%	Vision-Language Transformer
CLIPCap	MS COCO	~42%	Zero-shot capable
Your Implementation (EfficientNetB7 + LSTM + Attention)	Flickr8k	~9.71%	Comparable to early baselines

7.3) Summary

- While the paper's results are reasonable for the baseline model, they fall short of modern transformer-based approaches.
- The main contribution of the paper is hyperparameter analysis , not pushing the SOTA.
- My implementation matched the general methodology but lacks full convergence due to smaller dataset and fewer epochs.

8) What would you like to criticize about the paper? Could you suggest any improvements.

8.1) Criticisms

1. Limited Model Scope
   - Only evaluated the Show, Attend and Tell architecture
   - No comparison with more recent Vision Transformers or hybrid models
2. Single Dataset Evaluation
   - Focused solely on MS COCO
   - Lacks evaluation on specialized or low-resource datasets
3. Evaluation Metrics
   - Relied heavily on BLEU-4 and Top-5 Accuracy
   - Missed newer metrics like CIDEr , SPICE , or BERTScore
4. No Qualitative Analysis
   - No sample captions shown
   - No attention maps or visualization of errors
5. Hyperparameter Search Was Limited
   - Tested only a few combinations
   - No ablation study or detailed hyperparameter sweep

8.2) Suggested Improvements

IMPROVEMENT	BENEFIT
Add Vision Transformers (ViT, DeiT)	Better performance and generalization
Include qualitative examples	Show generated captions and attention maps
Add beam search decoding	Higher-quality captions
Report additional metrics (CIDEr, METEOR)	Better correlation with human judgment
Train on larger datasets	Improve robustness and generalization
Add metadata integration	Improve real-world applicability

By implementing these changes, the paper could become more practical for real-world deployment and aligned with current trends.

9) Have you implemented the paper using your own code? Do your results agree with the authors? What are the differences and why?

9.1) Have you implemented the paper using your own code?

Yes, I have implemented a simplified version of the paper using my own code:

- Used EfficientNetB7 as the image encoder
- Built a custom LSTM-based decoder with attention
- Trained on the Flickr8k dataset (publicly accessible alternative to MS COCO)
- Used teacher forcing, cross-entropy loss , and Adam optimizer , following the paper's setup

9.2) Do My Results Agree with the Authors?

ASPECT	PAPER'S RESULT	MY IMPLEMENTATION	AGREEMENT
Model Architecture	Show, Attend and Tell + ResNet/ VGG/ResNext	LSTM-based Decoder + EfficientNet Encoder	Similar approach
Loss Function	Cross-Entropy Best Performing	Used Cross-Entropy	✅ Matches
Optimizer	Adam performed best	Used Adam	✅ Matches
BLEU-4 Score	Not explicitly reported	~9.71%	LOWER
Attention Visualization	Qualitative examples shown	No visualization added yet	❌ Not Replicated
Training Duration	Multiple epochs with tuning	Limited training	❌ Shorter Training
Hyperparameter Selection	Extensive comparison done	Basic setup only	❌ Simplified

So while the core concepts and model structure align well, the quantitative performance metrics (like BLEU-4 score) cannot be directly compared at this stage due to:

- Limited number of training epochs
- Use of greedy decoding instead of beam search
- Use of a smaller public dataset (Flickr8k vs original private dataset)

9.3) Differences and Why They Exist

DIFFERENCE	REASON
Smaller Dataset	Flickr8k has fewer images → worse generalization
Shorter Training	Only 20 epochs → model may not have converged
Greedy Decoding	Beam search wasn't used → lower fluency and diversity
No Attention Visualization	Misses insight into where the model focuses
No Beam Search or SCST	Limits final performance
Different Encoder (EfficientNet instead of ResNet/VGG)	May affect feature alignment

Despite these differences, the core concepts — attention, teacher forcing, Adam optimization, and cross-entropy loss — were successfully replicated , making it a valuable educational and experimental exercise.

9.4) Conclusion

Yes, I have successfully implemented the core components of the paper using my own code, including:

- Feature extraction from images using pretrained CNNs
- Text preprocessing with tokenization
- Custom data generator
- LSTM-based decoder with teacher forcing
- Training loop and evaluation framework

While my implementation follows the same methodology and principles, there are differences in performance and scope due to:

- Dataset size and availability
- Training duration
- Absence of advanced decoding techniques like beam search
- Use of a different encoder (EfficientNet instead of VGG/ResNet)

10) Github Link

You can find our Github repository on the following link:

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

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/

[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