Exploring Image Captioning with Multiple Neural Network Models - Deep Machine Learning

Pasquale Bianco, Mattia Carlino

Chalmers University of Technology
Electrical Engineering Department

CHALMERS UNIVERSITY OF TECHNOLOGY

Introduction

In this project, we explore **Image Captioning** using multiple NN models; we use the **Flickr30K dataset**, which contains 30,000 images, each associated with 5 human–annotated captions, allowing more flexibility both in final evaluation and in loss for back–propagation stage.

Model Architectures Overview

- **ResNet50 + RNN:** uses ResNet50 as the CNN encoder to extract image features and LSTM RNN for caption generation. This is the simplest model, providing a baseline for comparison;
- **ResNet50 + RNN + Attention:** Incorporates an attention mechanism into the LSTM RNN architecture, allowing the model to focus on different regions of the image while generating captions, leading to improved performance on more complex images;

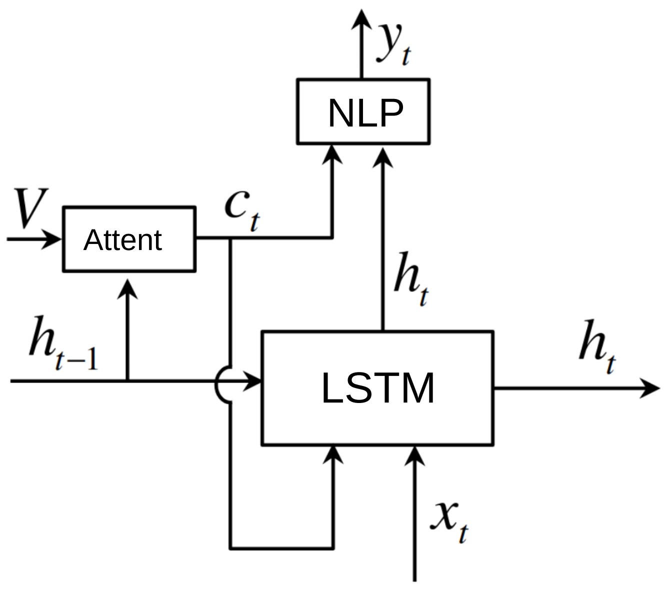

Figure 1: LSTM with Attention architecture

- **Vision Transformer (ViT) + RNN:** Leverages the Vision Transformer to extract global image features, which are then processed by the LSTM RNN for caption generation;
- **Microsoft GIT:** A pre-trained Generative Image-to-Text (GIT) model that uses CLIP tokens to generate captions, fine–tuned on current Flickr30k dataset. It employs Transformer-based decoding.

Model training losses

- **Attention mechanism** is able to better maintain information and details, but validation loss is almost the same with vanilla RNN;
- **ViT** is extracting better and more accurate features, allowing a more accurate caption generation;
- **GIT** is the best performing so far, achieving very low loss

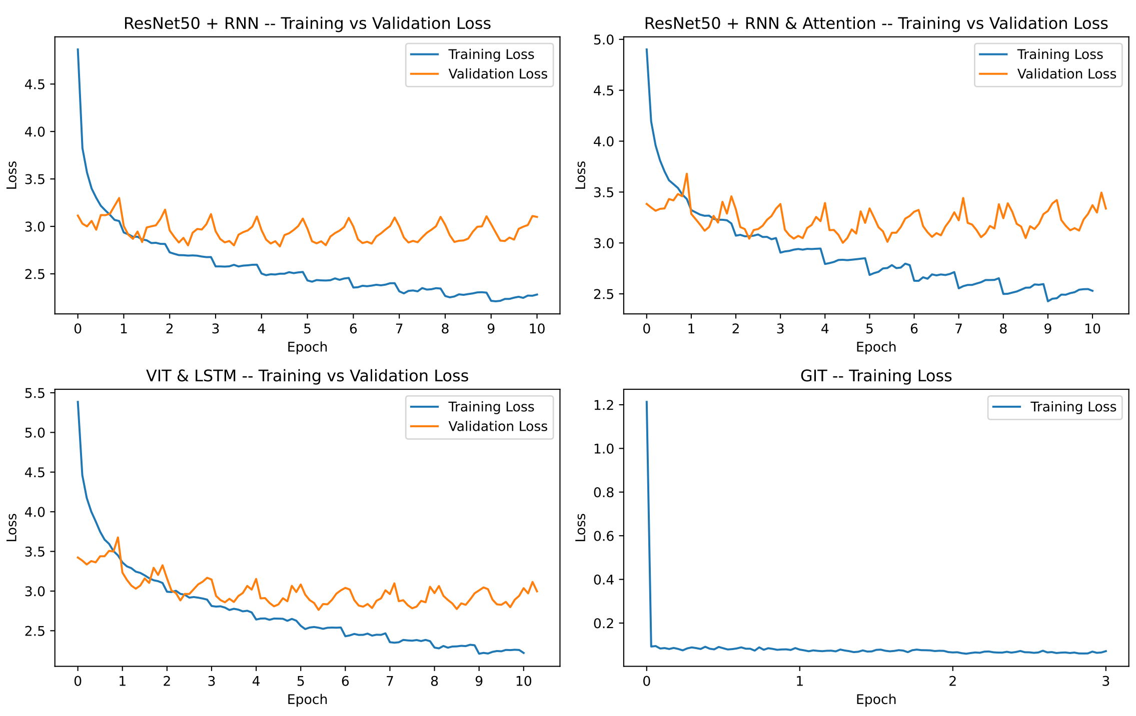

Figure 2: Training vs Validation loss in all models

Attention Mechanism

Attention mechanism allow the model to dynamically attends to different parts of the image and then use this information to generate meaningful and contextually relevant captions. Firstly, data from decoder and encoder are gathered through linear layers and activated with ReLU:

$$Attention = ReLU(W_E E + B_E \ + \ W_D D + B_D) \quad (1)$$

Then, this information can be handled to find separation hyperplanes:

$$Context = Softmax(Attention \cdot W_A + B_A) \quad (2)$$

Evaluation Metrics: BLEU and ROUGE

For evaluating the quality of the generated captions, we use two NLP metrics:

- **BLEU:** it measures the precision of n–grams[1] in *generated vs reference* captions, providing insight into the overlap of short sequences of words:
- **ROUGE:** it focuses on recall, assessing how well the generated captions capture important elements of the reference captions considering given embeddings.

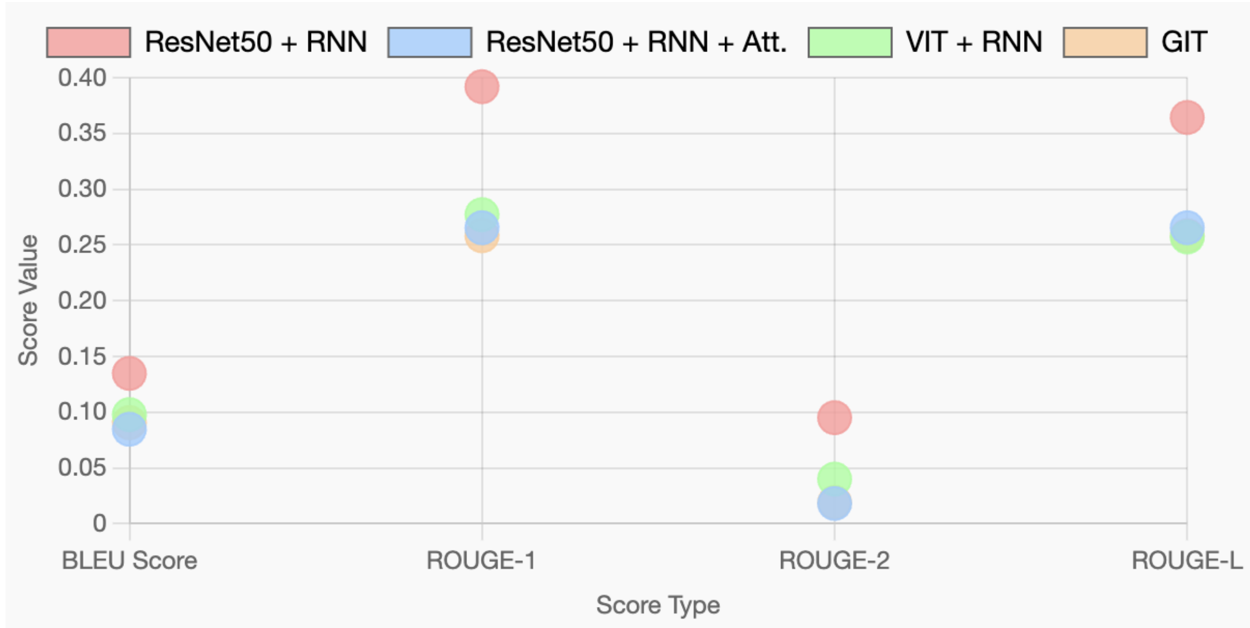

Figure 3: Performance comparison of the models using these metrics, related to Figure 4

[1] An n–gram is a contiguous sequence of *n* items from a given sample of text; items can be words, characters, etc..

Example of Caption Generation

Figure 4: Input Image

RN50 + RNN: *a snowboarder in a blue jacket and blue pants is skiing down a mountain*

RN50 + RNN & Att.: *a snowboarder is jumping over a snowy hill*

ViT + RNN: *snowboarder skiing down a snowy hill*

GIT: *a snowboarder jumps off a snowy hill*

Final Considerations

There is not a definitive best model:

- **RNN + Att.** and **GIT** achieve comparable scores (32.21% vs 37.70%) and together they are better in almost 70% of captions;
- **vanilla RNN** and **ViT + RNN**, instead, achieve worse results (17.26%, 12.83%), but still being better than previous models for some inputs.

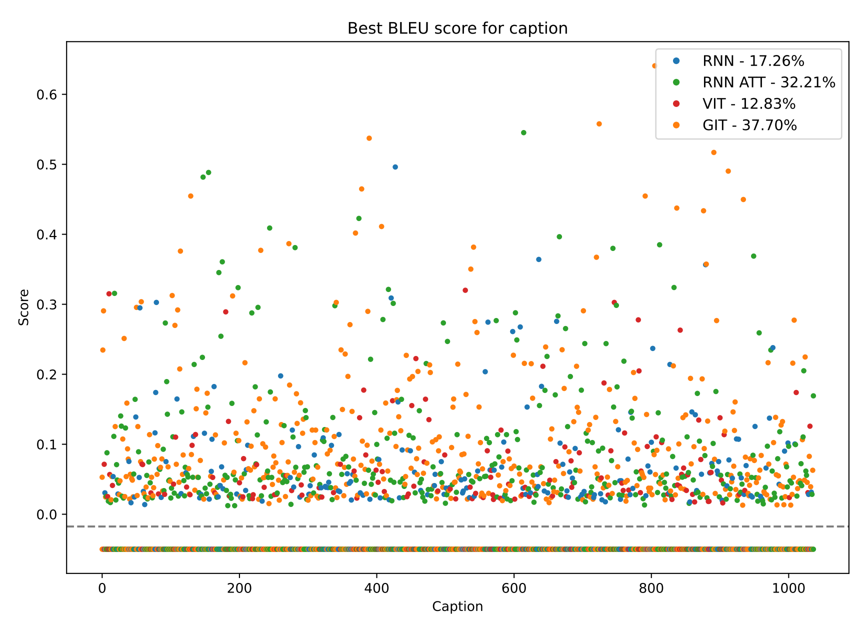

Figure 5: Best performing models on BLEU score

References

[1] S. Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy. Efficient image captioning based on vision transformer models. *Computers, Materials & Continua*, 73, 2022.

[2] M. H. J. H. Peter Young, Alice Lai. Flickr 30k dataset, 2014.

[3] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, Dec 2022.

[4] Z. Zohourianshahzadi and J. K. Kalita. Neural attention for image captioning: review of outstanding methods. *Artificial Intelligence Review*, 55(5):3833–3862, 2022.

October 28, 2024

pasquale@student.chalmers.se carlino@student.chalmers.se