Leveraging Early Exit and Rejection in ML for Efficiency

Florian Valade

Gustave Eiffel University, Paris

September 18, 2024

Table of Contents

1. Early Exit in Neural Networks
    - Introduction
    - ResNet Architecture
    - Motivation for Early Exiting
    - Early Exit Mechanisms
    - Benefits of Early Exiting
2. Early Exit in Large Language Models
    - Application to LLMs
    - Token Difficulty Identification
3. Conclusion

Introduction to Deep Neural Networks

- Neural networks process data through multiple layers, one after the other.
- Each layer extracts more complex features from the data.
- Deep networks can learn intricate patterns but may be computationally intensive.

ResNet: Deep Residual Networks

- ResNet (Residual Network) is a popular deep neural network architecture.
- Introduces **skip connections** to ease training of very deep networks.
- As data advances through layers, features become more refined and complex.

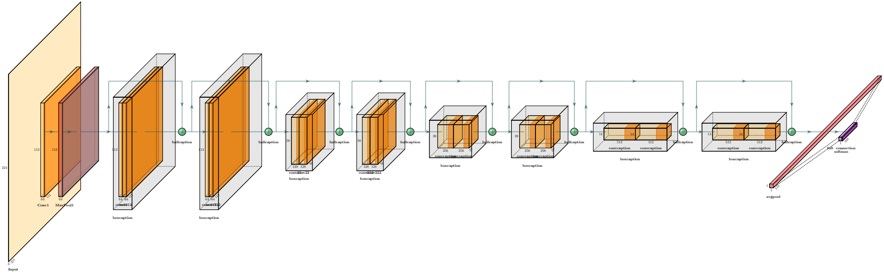
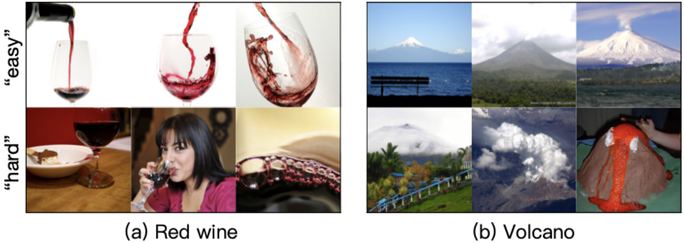


Figure 1. Structure of a ResNet with skip connections.

Motivation for Early Exiting

- Not all data requires processing through all layers.
- **Easy data**: Can be correctly predicted with fewer layers.
- Saves computational resources by not processing all data equally.



(a) Red wine

(b) Volcano

Figure 2. Some inputs are easier to classify than others. [1]

Images are from the ImageNet dataset.

Introducing Early Exit Mechanisms

- Add **early exit** points in the network with auxiliary classifiers.
- Allows the network to make predictions at intermediate layers.
- Stops computation early if the prediction is confident.

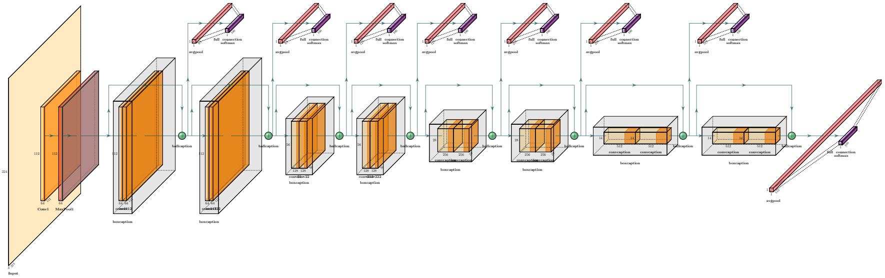


Figure 3. Neural network with early exit points.

Benefits of Early Exiting

- **Power Savings**: Less computation leads to reduced energy consumption.
- **Faster Inference**: Quick predictions for easy inputs.
- **Efficiency**: Optimizes resource usage.

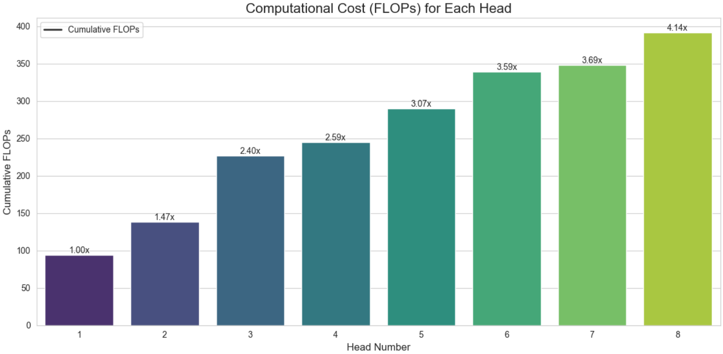


Figure 4. Difference of computation cost per head in ResNet with early exits.

Applying Early Exit to Large Language Models

- Large Language Models (LLMs) process sequences of words (tokens).
- **Some tokens are easier** to predict than others.
- Early exit can reduce computation in language tasks by exiting on easy tokens.

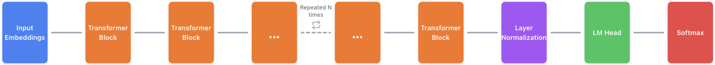
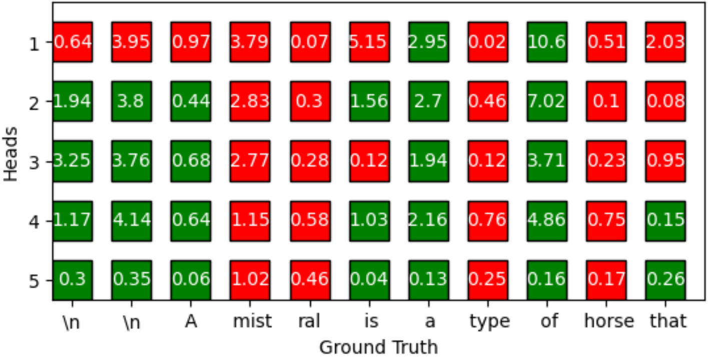


Figure 5. A simple LLM architecture diagram. The image shows a simple architecture with Transformer Blocks which are repeated N times.

Number of Transformer Blocks in Common LLMs		
GPT-2	:	12 to 48 blocks
GPT-3	:	96 blocks
LLaMA	:	32 to 80 blocks

Early Exit in Language Models: Identifying Token Difficulty

- Early exit helps identify easy and hard tokens in language processing.
- Allows for efficient allocation of computational resources.
- Improves processing speed for simpler parts of the text.



Heads	\n	\n	A	mist	ral	is	a	type	of	horse	that
1	🟥 0.64	🟥 3.95	🟥 0.97	🟥 3.79	🟥 0.07	🟥 5.15	🟩 2.95	🟥 0.02	🟩 10.6	🟥 0.51	🟥 2.03
2	🟩 1.94	🟩 3.8	🟩 0.44	🟥 2.83	🟥 0.3	🟩 1.56	🟩 2.7	🟥 0.46	🟩 7.02	🟥 0.1	🟥 0.08
3	🟩 3.25	🟩 3.76	🟩 0.68	🟥 2.77	🟥 0.28	🟥 0.12	🟩 1.94	🟥 0.12	🟩 3.71	🟥 0.23	🟥 0.95
4	🟩 1.17	🟩 4.14	🟩 0.64	🟥 1.15	🟥 0.58	🟩 1.03	🟩 2.16	🟥 0.76	🟩 4.86	🟥 0.75	🟩 0.15
5	🟩 0.3	🟩 0.35	🟩 0.06	🟥 1.02	🟥 0.46	🟩 0.04	🟩 0.13	🟥 0.25	🟩 0.16	🟥 0.17	🟩 0.26

Figure 6: Visualization of early exit identifying easy and hard tokens i

Conclusion

- Early exit optimizes neural network processing by leveraging data difficulty.
- Saves power and improves inference speed without sacrificing accuracy.
- Applicable to various models, enhancing both performance and efficiency.

Thank You!