SVD-LoRA: Singular Value Decomposition Compression of Low Rank Adaptations

Ali Hussein

Sunbird AI

Introduction

- Model sizes have escalated in recent years; bigger models often perform better.
- Scaling both training data and model parameters improves performance.
- Larger models limit AI research for those without massive computational resources.
- Existing strategies for scaling include quantization, pruning, and distillation.

Related Work

- *Quantization*: Reduces weight precision, but might decrease model performance.
- *Pruning*: Removes redundant connections; requires fine-tuning.
- *Low Rank Adaptation*: Represents weights in lower-dimensional space.
- *Distillation*: Trains a smaller model to mimic a larger one's outputs.

Methodology

- Use SVD to represent original neural network weights with low-rank approximations.
- Represent weight matrix with two separate low-rank matrices.
- Construct a difference matrix between original matrix and SVD. Use SVD to compress the difference matrix.
- The compressed difference matrices are added to the existing SVD low-rank matrices. (Inspired by the model saving practices of the Stable Diffusion community)
- This step is repeated until a required number of submatrices is generated, with each new difference matrix is the result of the reconstructed matrix (using all SVD steps) with the original matrix
- Approach reduces storage and computational complexity, while preserving model quality.
- Methodology can be generalized to use multiple matrices.

Experiments

- Tested on CIFAR-10 and NMT datasets.
- Different matrix counts and ranks were evaluated.
- Results vary but show potential.

Submatrices	Eval Accuracy	Eval Runtime
Baseline	**0.9769**	131.2141
1	0.9571	129.2873
2	0.962	**128.7298**
3	0.9617	129.0121
4	0.9622	129.2541
5	0.9619	129.3435
6	0.9611	129.3593
7	0.9588	129.4201
8	0.9613	129.3076
9	0.9612	130.1609

Table 1. CIFAR-10 accuracy across 1-9 sub-matrices with a fixed number of parameters

Parameter %	Eval Accuracy	Eval Runtime
Baseline	**0.9769**	131.2141
10%	0.9617	130.4624
20%	0.9746	**128.9813**
30%	0.9766	129.0198
40%	0.9763	129.067
50%	0.9767	129.135
60%	0.9767	129.3274
70%	0.9768	129.5422
80%	0.9764	129.5246
90%	0.9759	129.4248
100%	0.9759	129.7098

Table 2. CIFAR-10 Accuracy across 10-100% parameters at 3 sub-matrices

Factor	Mean	
	BLEU	chrf
Baseline	30.0678	58.495
30% (r=102)	19.8637	41.0294
40% (r=136)	**31.0818**	55.9282
50% (r=170)	30.437	**58.7765**
60% (r=204)	29.8488	58.4441
70% (r=238)	28.7292	57.9683
80% (r=272)	28.392	57.7862
90% (r=306)	29.4163	58.3949
100% (r=340)	29.6074	58.3529

Table 3. Mean (5 salt language) Evaluation Metrics at different parameter counts

Unsuccessful Directions

- Attempts to directly parameterize a randomly initalized lower rank matrices with backpropagation or evolutionary optimization were unsuccessful.
- A loss of mean squared error between reconstructed matrix and original matrix as well as difference between norms and difference between singular values did not lead to convergence

Conclusion

- Introduced SVDLoRA method for model compression.
- Demonstrated its potential for model robustness and usability.

Future Work

- Further studies are required.
- Potential application with Switch Transformers and Mixture of Experts models.
- Compare SVDLoRA with other compression methods.
- Theoretical explorations into the lower bound of compression for neural network weights.