Attention-Informed Optimization: Leveraging Attention Energies for Neural Network Training

Jeronimo Aranda Barois
December 9, 2024

Abstract

We present a novel optimization approach for transformer-based neural networks that leverages the inherent energy landscape computed by attention mechanisms during forward passes. Unlike previous approaches that require additional computation to estimate energy landscapes, our method utilizes the already-computed attention energies to inform parameter updates. We show that this attention-informed optimization can lead to more efficient training by naturally identifying important parameter regions through the model's own attention patterns.

1. Introduction

The optimization of deep neural networks remains a crucial challenge in machine learning. While methods like Adam have become standard, they do not directly utilize the rich structural information encoded in attention mechanisms. This paper introduces a new optimization approach that bridges this gap by incorporating attention energies into the parameter update process.

2. Related Work

Previous work has explored various approaches to neural network optimization:

- Adaptive methods like Adam (Kingma & Ba, 2014)
- Energy-based optimization approaches
- Attention mechanisms in transformers (Vaswani et al., 2017) However, none have directly utilized attention energies for optimization.

3. Method

3.1 Attention Energy Landscape

In transformer architectures, the attention mechanism computes energy scores:

```
E(Q,K) = QK^T/√d
```

These scores, before softmax normalization, provide a natural energy landscape across token relationships. We propose that these energies contain valuable information about the parameter space.

3.2 Attention-Informed Updates

We modify the Adam optimizer to incorporate attention information:

```
def attention_informed_update(params, grad, attention_energies, β):
    # Standard Adam components
    m = β₁ * m + (1 - β₁) * grad
    v = β₂ * v + (1 - β₂) * grad²

    # Attention energy gradient
    attention_grad = compute_attention_gradient(attention_energies)

    # Combined update
    update = α * (m / (√v + ε)) + (1 - α) * attention_grad

    return update
```

Where α is dynamically adjusted based on the attention pattern strength:

```
α = sigmoid(γ * mean(attention_energies))
```

3.3 Theoretical Analysis

The attention energies provide several beneficial properties:

1. Natural parameter importance weighting
2. Multi-scale optimization landscape information
3. Implicit regularization through attention patterns

4. Experiments

We evaluate our method on several standard benchmarks:

1. Language modeling tasks
2. Vision transformer training

3. Multi-modal learning

Metrics include:

- Training convergence speed
- Final model performance
- Computational overhead
- Attention pattern stability

5. Results

[Placeholder for experimental results]

6. Discussion

The proposed method offers several advantages:

1. Zero additional forward passes required
2. Natural incorporation of model structure
3. Adaptive optimization based on attention patterns

7. Conclusion

We present a novel optimization approach that leverages attention energies to guide parameter updates. This method provides a more informed optimization process without additional computational overhead.

Acknowledgments

Special thanks to Rodolfo Arturo, Javier D, and all friends, family, and people who supported this work along the way.

References

1. Vaswani, A., et al. (2017). Attention is all you need.
2. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization.
3. [Additional relevant references]