1

Introduction

In deep learning, normalization techniques are essential for training deep neural networks efficiently and effectively. Normalization helps stabilize the learning process and accelerate the training of deep models by ensuring that the input features are on a similar scale. Various normalization methods have been proposed over the years, each with its strengths and weaknesses. This report delves into Batch Normalization (BN), its problems, and the evolution of other normalization techniques like Layer Normalization (LN), Instance Normalization (IN), and finally Group Normalization (GN), which aims to combine the best of both LN and IN.

1.1 Batch Normalization (BN)

Batch Normalization was introduced as a method to normalize the inputs of each layer so that they have a mean of zero and a variance of one. This technique helps mitigate the issue of internal covariate shift, where the distribution of inputs to a learning model changes over time as the parameters of the previous layers change.

1.1.1 Problems with Batch Normalization

- **Dependency on Batch Size**: BN's effectiveness is highly dependent on the batch size. When the batch size is too small, the estimated statistics (mean and variance) become unreliable, leading to degraded performance.
- **Inconsistency During Training and Inference**: During inference, the batch statistics are replaced with running averages computed during training, which might not always accurately represent the test data distribution.
- **Limited Use in Certain Applications**: Applications like object detection, segmentation, and video processing often require small batch sizes due to memory constraints, where BN performs poorly [**?**].

1.2 Layer Normalization (LN)

Layer Normalization was introduced to address the drawbacks of BN. LN normalizes across the features within each layer for each training example, rather than across the batch.

1.2.1 Advantages of Layer Normalization

- **Independence from Batch Size**: LN computes the normalization statistics across the features of each layer independently for each sample, making it suitable for models with varying batch sizes or even batch size of one.
- **Effective for Sequential Models**: LN has been particularly effective in training sequential models like RNNs and LSTMs, where the batch dimension is not as well-defined.

1.2.2 Drawbacks of Layer Normalization

- **Limited Performance in Vision Tasks**: Although LN works well in sequential tasks, it has not achieved the same level of performance as BN in visual recognition tasks [?].

1.3 Instance Normalization (IN)

Instance Normalization is another technique that extends the idea of normalization by normalizing each instance in the batch independently. This method is particularly popular in style transfer applications.

1.3.1 Advantages of Instance Normalization

- **Per-Instance Normalization**: By normalizing each instance independently, IN effectively handles style variations, making it suitable for generative tasks like style transfer.

1.3.2 Drawbacks of Instance Normalization

- **Suboptimal for Discriminative Tasks**: IN does not perform as well as BN in discriminative tasks like classification and detection because it lacks the ability to normalize across the batch dimension, which can help stabilize the learning process for these tasks [?].

1.4 Group Normalization (GN)

Group Normalization is introduced as a solution that combines the advantages of both Layer Normalization and Instance Normalization. GN divides the channels into groups and computes the normalization statistics within each group.

1.4.1 Advantages of Group Normalization

- **Stable Across Different Batch Sizes**: GN's computation is independent of batch size, which makes it stable and effective even with very small batches or a batch size of one.
- **Combines Strengths of LN and IN**: GN leverages the structure of LN by normalizing across channels and the flexibility of IN by grouping channels, leading to improved performance in a variety of tasks.

1.4.2 Drawbacks of Group Normalization

- **Complexity in Tuning Group Size**: The performance of GN can be sensitive to the choice of the number of groups, requiring careful tuning to achieve optimal results [**?**].

1.5 Motivation

Batch Normalization has significantly advanced deep learning by normalizing the mean and variance of features within a batch, facilitating smoother optimization and faster convergence. However, BN's effectiveness diminishes with small batch sizes due to inaccurate batch statistics estimation, leading to increased error rates. This limitation is particularly critical in tasks like object detection, segmentation, and video classification, where high-resolution images necessitate smaller batch sizes. Group Normalization, introduced by Yuxin Wu and Kaiming He from Facebook AI Research, addresses this issue by dividing channels into groups and normalizing the features within each group. GN's computation is independent of batch sizes, maintaining stable accuracy across varying batch sizes and providing consistent performance from pre-training to fine-tuning stages.