Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th

Deep Learning

Deep Learning (Chambers and Zaharia, 2018)

- Deep learning is a rapidly evolving area in Spark, particularly effective in handling unstructured data such as images, audio, and text.
- Deep learning involves neural networks, which are layers of nodes with weights and activation functions stacked together.
- As data passes through these layers, more complex features are recognised.

Deep Learning (Chambers and Zaharia, 2018)

- Advances in large datasets, hardware, and training algorithms have made deep learning dominant in fields like computer vision and speech processing.
- Spark's parallel computing framework complements deep learning well, enabling efficient processing of large datasets.

Deep Learning on Spark

(Chambers and Zaharia, 2018)

- Inference: Applying pretrained models to large datasets using Spark, such as image classification models, allows parallelised processing.
- Featurization and Transfer Learning: Transfer learning uses features from pretrained models to solve new problems, which is especially useful when training data is limited.
- Model Training: Spark enables parallel training of models, either by distributing training across clusters or running multiple models for hyperparameter tuning.

Deep Learning on Spark

(Chambers and Zaharia, 2018)

- TensorFlowOnSpark is a popular library that facilitates parallel training of TensorFlow models on Spark clusters.
- While TensorFlow supports distributed training, it lacks a built-in cluster manager and distributed I/O capabilities.

Deep Learning on Spark

(Chambers and Zaharia, 2018)

- TensorFlowOnSpark addresses this by launching TensorFlow's distributed mode within a Spark job and feeding data from Spark RDDs or DataFrames directly into the TensorFlow job.
- This integration allows users familiar with TensorFlow's distributed mode to easily run jobs on Spark clusters and leverage Spark's data processing capabilities.

Deep Learning on Spark

(Chambers and Zaharia, 2018)

- Deep Learning Pipelines is an open-source package from Databricks that integrates deep learning into Spark's ML Pipelines API, focusing on ease of use and automatic distribution of computation.
- It supports frameworks like TensorFlow and Keras, incorporating them into standard Spark APIs (e.g., ML Pipelines and Spark SQL).

Deep Learning on Spark

(Chambers and Zaharia, 2018)

- The package simplifies tasks like building transfer learning models with the DeepImageFeaturizer class and allows for parallel grid search and cross-validation using MLlib's APIs.
- Additionally, users can export models as Spark SQL user-defined functions for SQL or streaming applications.
- The package is actively being developed, with ongoing updates.

Test2
(5 points)

References

- Chambers, B., & Zaharia, M. (2018). Spark: The definitive guide: Big data processing made simple. " O'Reilly Media, Inc.".