11-792 Project Report

Nicholas Gekakis, Boyue Li

January 29, 2018

1 Overview

In this project, we are building a distributed question answering pipeline framework, which allows users to easily configure multiple modules, create complex pipelines and tune parameters automatically.

2 Requirements

2.1 Easy to configure and deploy

The framework should be easy to configure and deploy.

2.2 Save and resume

The framework should be able to save intermediate results to resume training.

2.3 Pipeline topology

The framework should be able to create pipelines with complex topologies.

2.4 Automatically parameters tuning

The framework should be able to automatically tune some parameters.

2.5 Distributed parallel processing

The framework should support distributed parallel processing to handle large datasets and complicated pipelines.

3 Design

A pipeline is constructed by several independent modules, users only need to specify the correspondance between inputs and outputs, the framework would automatically handle all calculation.

3.1 Module

A module is the basic calculation unit which takes arbitrary number of inputs and produces arbitrary number of outputs. Every input and output is an information object defined below. A module also needs a description file to specify following fileds:

- Number of inputs.
- Number of outputs.
- Data type of inputs.
- Data type of outputs.
- Number of parameters.
- Default values of parameters.
- Tuning interval of parameters.
- Tuning steps of parameters.

3.2 Information object

The information obejct used to pass data between modules contains following fileds:

- Producing module: the module that produced this information object.
- Consuming module: the module that this information object to be passed to.
- Data path: the path to actual data file.
- Data type: the type of data (one of number, binary and string or user defined data type).
- Data size: the number of data instances.

3.3 File format

A data file has a description file which contains following fileds:

- The configuration applied to it.
- The timestamp when it is created.
- Data type.
- Data size.

3.4 Configuration file

We use YAML files to configure the framework.

3.5 Code structure

4 Experiments

4.1 Dataset1

4.2 Dataset2

4.3 Dataset3

5 Conclusion

Acknowledgements