Keras: Performance Analysis of Tensorflow, Theano, and CNTK Backends

R244 Presentation By: Vikash Singh  November 28, 2018 Session 8
What is Keras[^1]?

- High level neural networks API in Python that is capable of running on top of TensorFlow[^2], CNTK[^3], or Theano[^4]
- Focus on enabling fast experimentation
- Supports feedforward, convolutional, and recurrent neural networks
- Runs on both CPU and GPU
Code Example: Understanding Keras Abstractions

```python
model = Sequential()
model.add(Dense(512, activation='relu', input_shape=(784,)))
model.add(Dropout(0.2))
model.add(Dense(512, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(num_classes, activation='softmax'))

model.summary()

model.compile(loss='categorical_crossentropy',
              optimizer=RMSprop(),
              metrics=['accuracy'])

history = model.fit(x_train, y_train,
                     batch_size=batch_size,
                     epochs=epochs,
                     verbose=1,
                     validation_data=(x_test, y_test))

score = model.evaluate(x_test, y_test, verbose=0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
```
Analyzing Performance with Different Backends

- “Backend engine” of Keras takes high level building blocks and converts to low level code
- **Goal:** Experimentally analyze the performance of different backends in different computing environments (local vs cluster)
- Practical usage for understanding if there is a significant difference, can be exploited in production use case
Experimental Methods

● Train deep neural network locally (CPU) on MNIST digit recognition dataset using Keras and use Theano, TensorFlow, and CNTK backends

● Use Google Cloud ML to run the exact same job on three GPU environments:
  ○ A single NVIDIA Tesla K80 GPU
  ○ Four NVIDIA Tesla K80 GPUs
  ○ Eight NVIDIA Tesla K80 GPUs
Key Questions

- Are there any significant performance differences across the board?
- Do certain backends perform better in certain computing environments?
- Do we see identical model performance across the different backends?
- Can we link the design of these systems to the observed results?
Future Work

- Analyzing these trends not only across various computing environments but also different neural network models (CNNs, RNNs, etc)
- Testing on more heterogeneous computing environments (mixtures of CPUs, GPUs, etc)
- Comparing Keras implementation to native implementation of exact same architectures
Progress

● Planned out experiments, selected platforms for running GPU cluster on cloud
● Began training neural networks on Keras locally using MNIST dataset
● To Do: Perform actual experiments (locally and using Google Cloud) and record/analyze results, see what parts of future work can be done
References