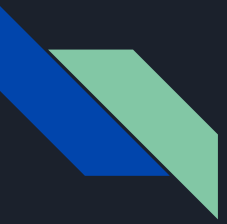
Enabling Polyhedral optimizations in TensorFlow through Polly

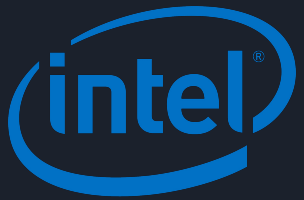

annanay25


cs14btech11001@iith.ac.in

Annanay Agarwal, IITH
Michael Kruse, Polly Labs
Brian Retford, Vertex.ai
Tobias Grosser, ETHZ, Polly Labs
Ramakrishna Upadrasta, IITH

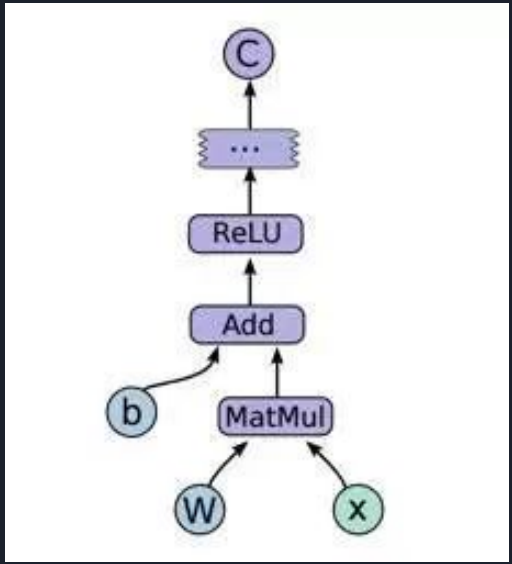


Fast Machine Learning for everyone.



A wild TensorFlow appears!

- TensorFlow - Open Source Deep learning Framework by Google.
- Built-in cross platform support for writing Machine learning code.
- Numerical computation using data flow graphs.



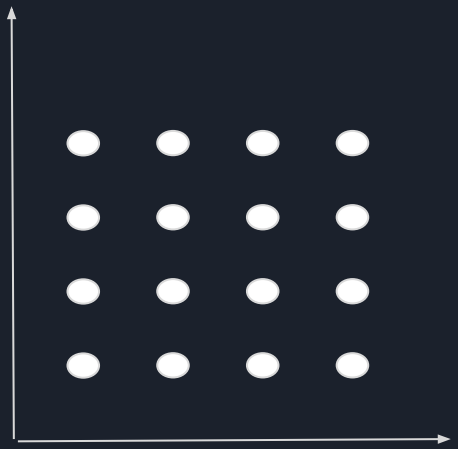
XLA - X(Acc)elerated Linear Algebra

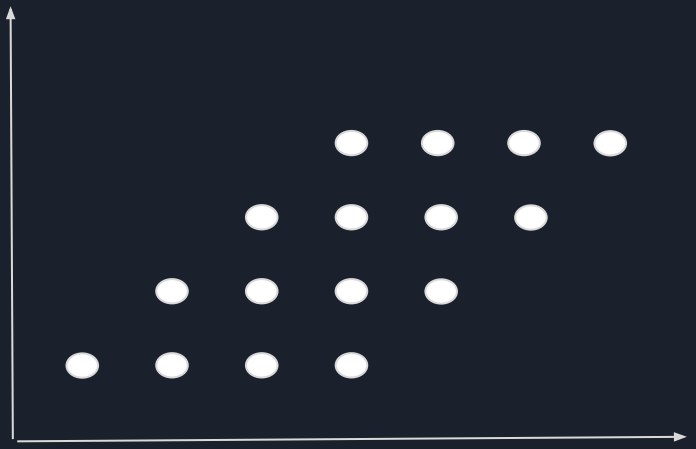
- Recently open sourced (Jan'17)!
- JIT (Just In Time) -
  - Runtime Compilation!
  - Know the size of dataset you are dealing with!
  - JIT compiles subgraphs of the TensorFlow computation!
- Uses LLVM as a backend!

The Force Awakens.

Architecture Diagram.

Polyhedral Compilation



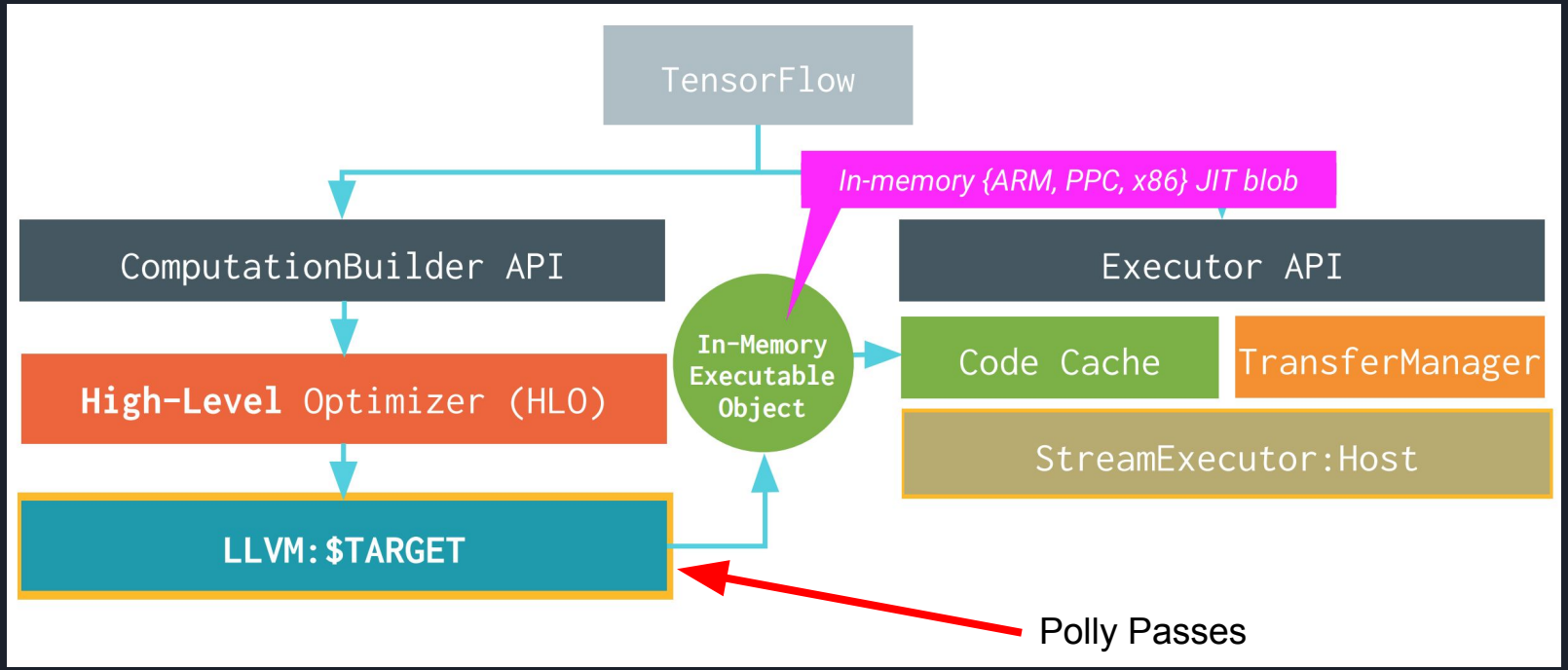


```
for ( i = 0; i < N; i++ )
  for ( j = 0; j < M; j++ )
    Stmt ( i, j )
```

```
for ( i = 0; i < N; i++ )
  for ( j = i ; j < M + i ; j++ )
    Stmt ( i, j - i )
```

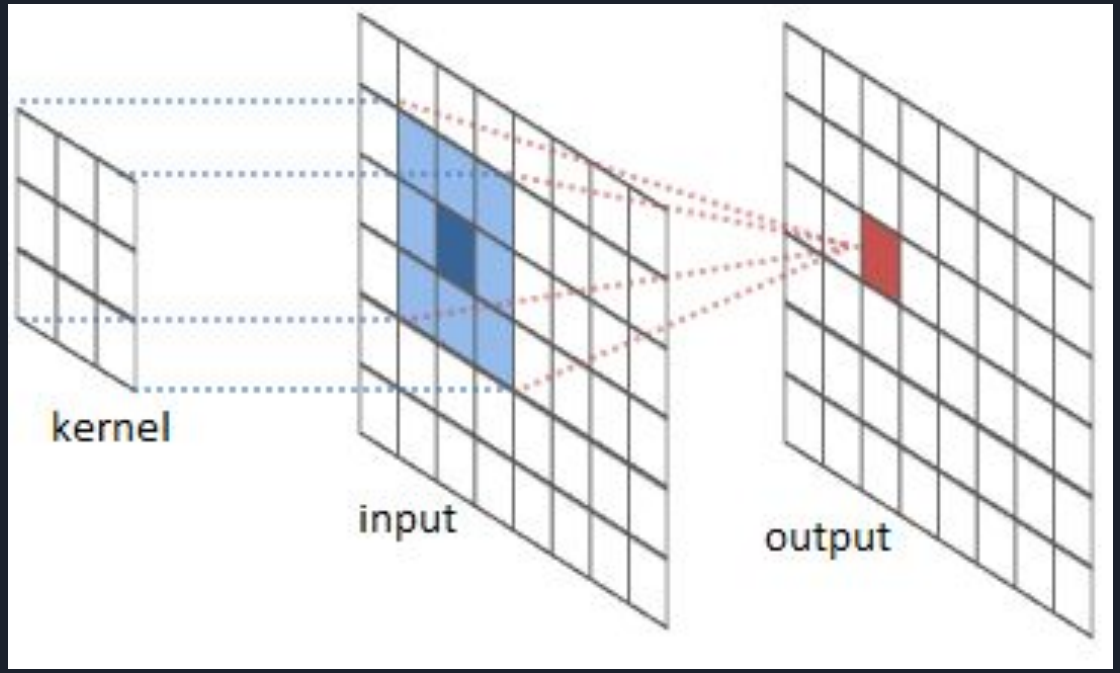
Pollyyy!

Polly in TensorFlow.



Behold!

Convolutional Neural Networks

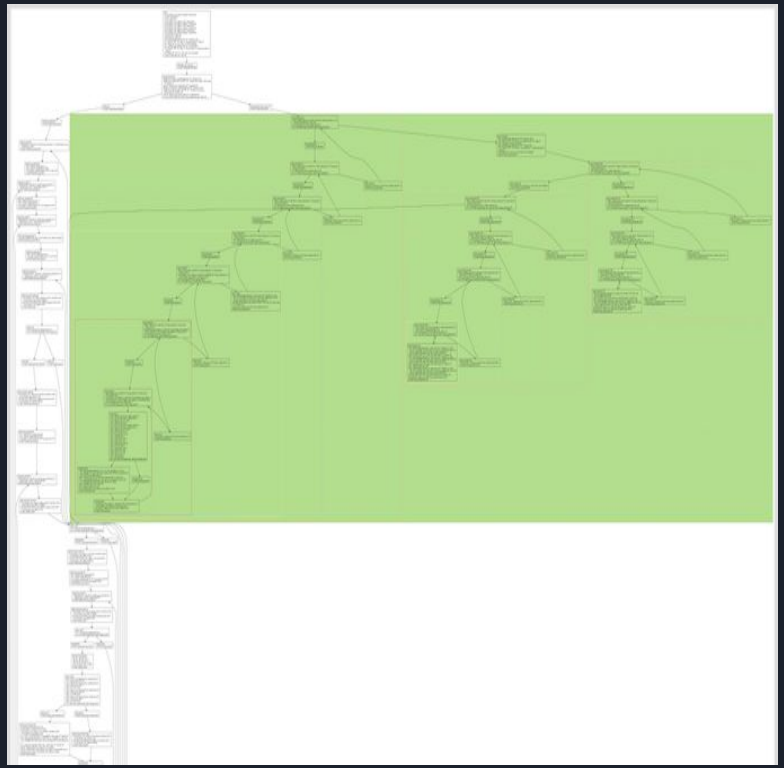



```
/* Conv 2D. */
for (int m = 0 ; m < 64 ; m ++ ){
  for (int i = 1; i < 31 ; i ++ ){
    for (int j = 1; j < 31 ; j ++ ){
      for (int k = 0; k < 3 ; k ++ ){
        for (int l = 0; l < 3 ; l ++ ){
          int temp1 = (img[i - 1 + k][j - 1 + l]);
          int temp2 = (kernel[k][l][m]);
          sum += temp1 * temp2;
        }
      }
      res[i-1][j-1][m] = sum;
      sum = 0;
    }
  }
}
```

Results

SCoP Detection

Polly's SCoP detection was modified to detect the convolution kernel from the LLVM IR generated for tf.conv2d() operation in XLA.

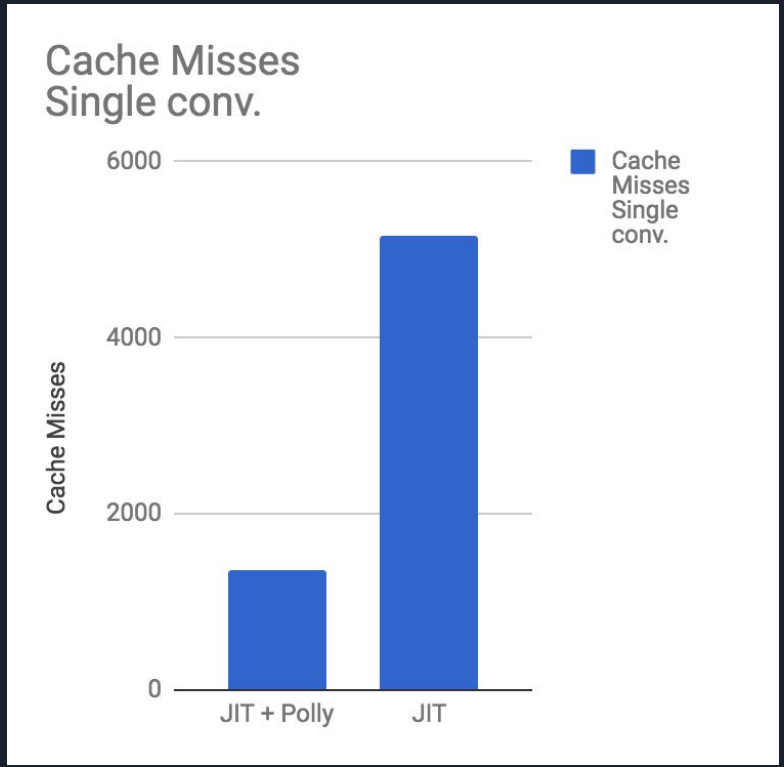


Results

Performance

Reduced cache misses - advanced data locality optimizations like tiling.

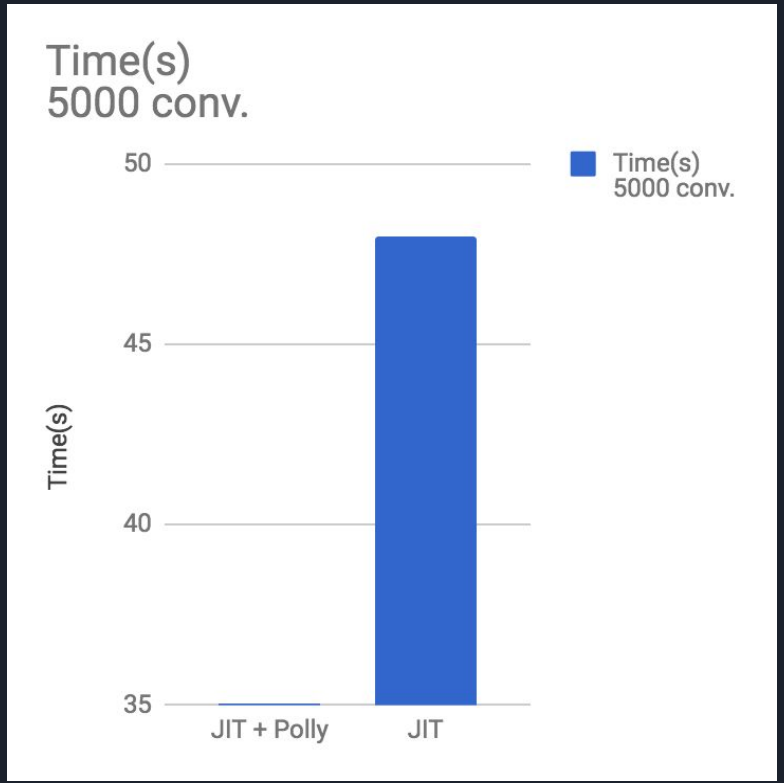
Also performs operator fusion.



Results

Performance

Better overall runtime for convolution kernel despite having a greater compilation time.



Future Work

- SCoP detection and pattern optimization does not work for other deep learning kernels like Recurrent Neural Networks (RNNs).
  - Expand support to more Deep Learning kernels.
- Polly is capable of generating GPGPU code.
  - Polly as a backend.

Thank you!

annanay25 cs14btech11001@iith.ac.in