v1.0 has been released!

- Steady grow of community
  - 70 contributors added 300K+ lines of codes

- AWS released ~10 ML products last week, all of them based on MXNet.

Monthly installation over last 6mons

<table>
<thead>
<tr>
<th></th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sep</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>0</td>
<td>600000</td>
<td>1200000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SageMaker
Hosted pipeline for training and deploy

Rekognition
Image and video analysis

Translate
Machine translation
Ease of use is No. 1 Target

- Mixed interface
  - **Performance**: Symbolic model definition
  - **Flexibility**: imperative tensor operation
  - **Portability**: several PL inference, Python, Scala, Perl, R, ...

- Our observations
  - 90% users are **new**, they want to get started quickly (in a few hours)
  - New DL models are **more structured**, more than chaining conv and fc layers
Gluon: a new imperative interface

- Collaborated between Amazon and Microsoft
- Inspired from others

```python
class MLP(gluon.Block):
    def __init__(self):
        super(MLP, self).__init__()

        with self.name_scope():
            self.dense0 = gluon.nn.Dense(64)
            self.dense1 = gluon.nn.Dense(64)
            self.dense2 = gluon.nn.Dense(10)

    def forward(self, x):
        x = nd.relu(self.dense0(x))
        x = nd.relu(self.dense1(x))
        x = self.dense2(x)
        return x
```

No need to give input size
Hybridize (JIT)

Pros:
- Switch to fast symbolic execution
- Portability

Cons:
- Doesn’t support dynamic programs
- EC2 P3.8xlarge: 4 Tesla Volta
- Resnet 50 training
Folding

[Figure adapted from Moshe Looks]

[Looks, et.al 2017]
Folding Performance

- EC2 C4.8xlarge
- Tree LSTM

Bar chart showing:
- # images / sec vs # GPUs
- Training and Inference performance
- w/o fold, w/o fold, hybridized, w/ fold
Summary

- Gluon: a new imperative interface
- Improve performance
  - Hybridizing
  - Folding