

MXNet Updates 2017

Junyuan Xie



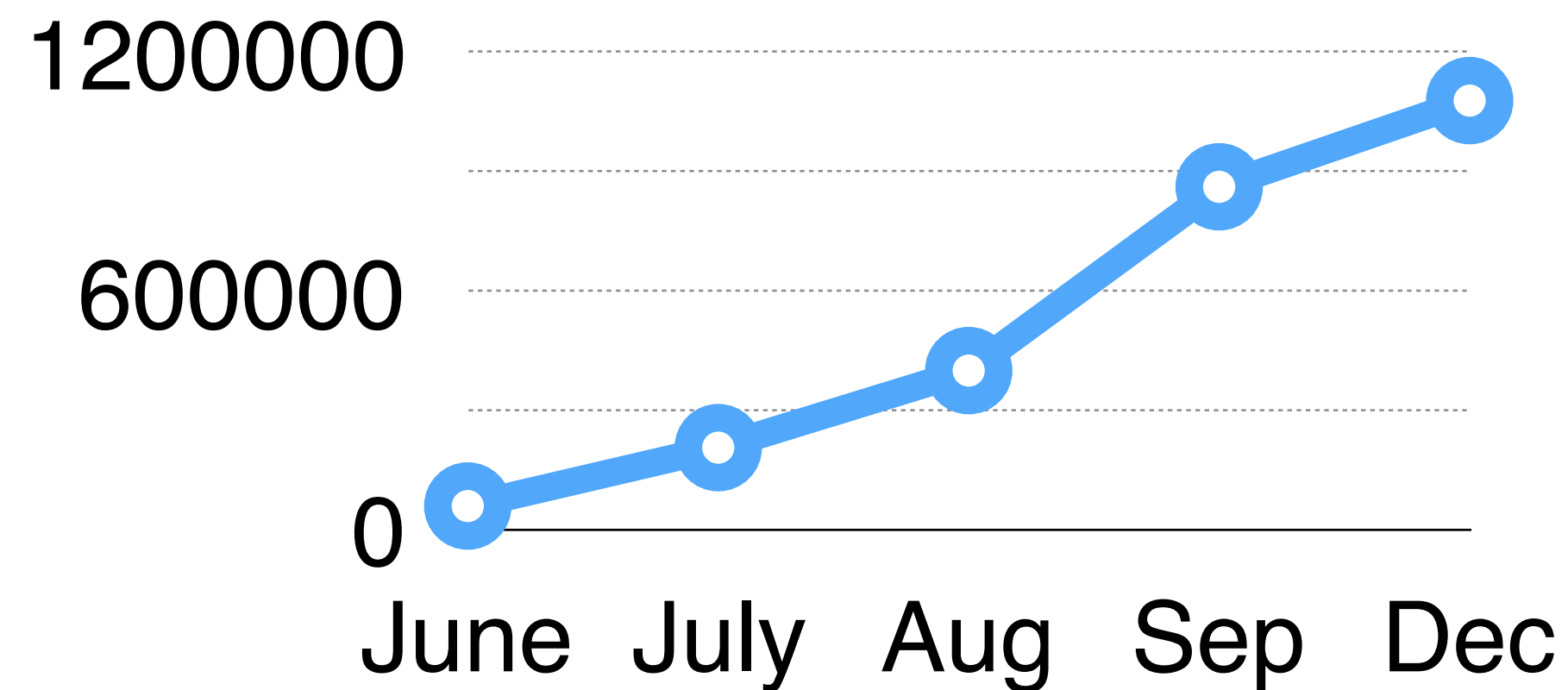
v1.0 has been released!

- ◆ 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



SageMaker

Hosted pipeline for training and deploy

Rekognition

Image and video analysis

Translate

Machine translation

Ease of use is No. 1 Target

◆ Mixed interface

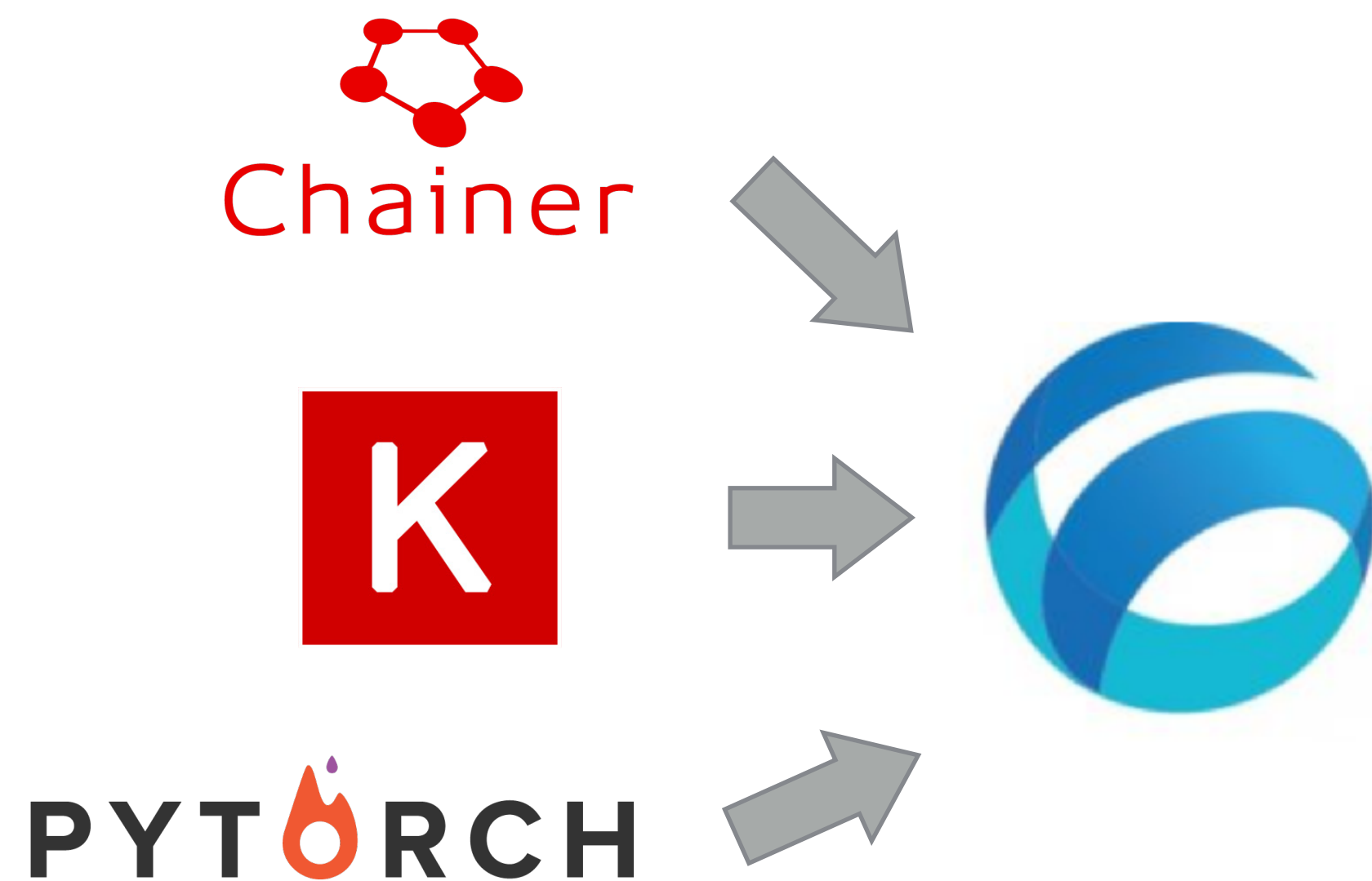
- ❖ **Performance**: Symbolic model definition
- ❖ **Flexibility**: imperative tensor operation
- ❖ **Portability**: several PL inference, Python, Scala, Perl, R, ...

◆ 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

Glueon: a new imperative interface

- ◆ Collaborated between Amazon and Microsoft
- ◆ Inspired from others



```
1 class MLP(gluon.Block):
2     def __init__(
3         super(MLP, self)
4         with self:
5             self.dense0 = gluon.nn.Dense(64)
6             self.dense1 = gluon.nn.Dense(64)
7             self.dense2 = gluon.nn.Dense(10)
8
9     def forward(self, x):
10        x = nd.relu(self.dense0(x))
11        x = nd.relu(self.dense1(x))
12        x = self.dense2(x)
13        return x
```

No need to give input size

Hybridize (JIT)

```
1 class Net(HybridBlock):
2     def __init__(self, **kwargs):
3         super(Net, self).__init__(**kwargs)
4
5         self.fc1 = nn.Linear(256)
6         self.fc2 = nn.Linear(128)
7         self.fc3 = nn.Dense(2)
8
9     def hybrid_forward(self, F, x):
10        x = F.relu(self.fc1(x))
11        x = F.relu(self.fc2(x))
12        return self.fc3(x)
```

Switch between imperative
and symbolic

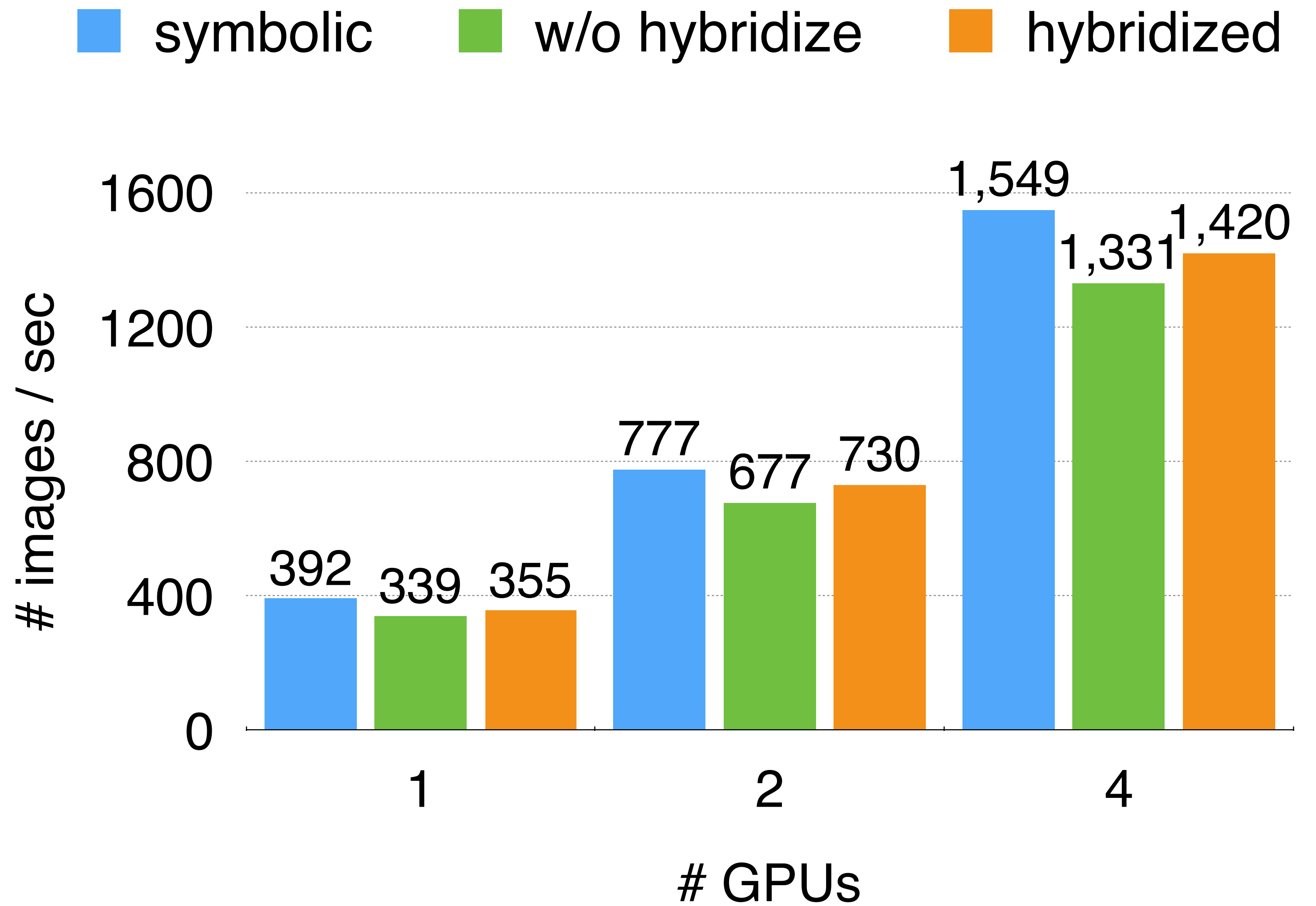
◆ Pros:

- ❖ Switch to fast symbolic execution
- ❖ Portability

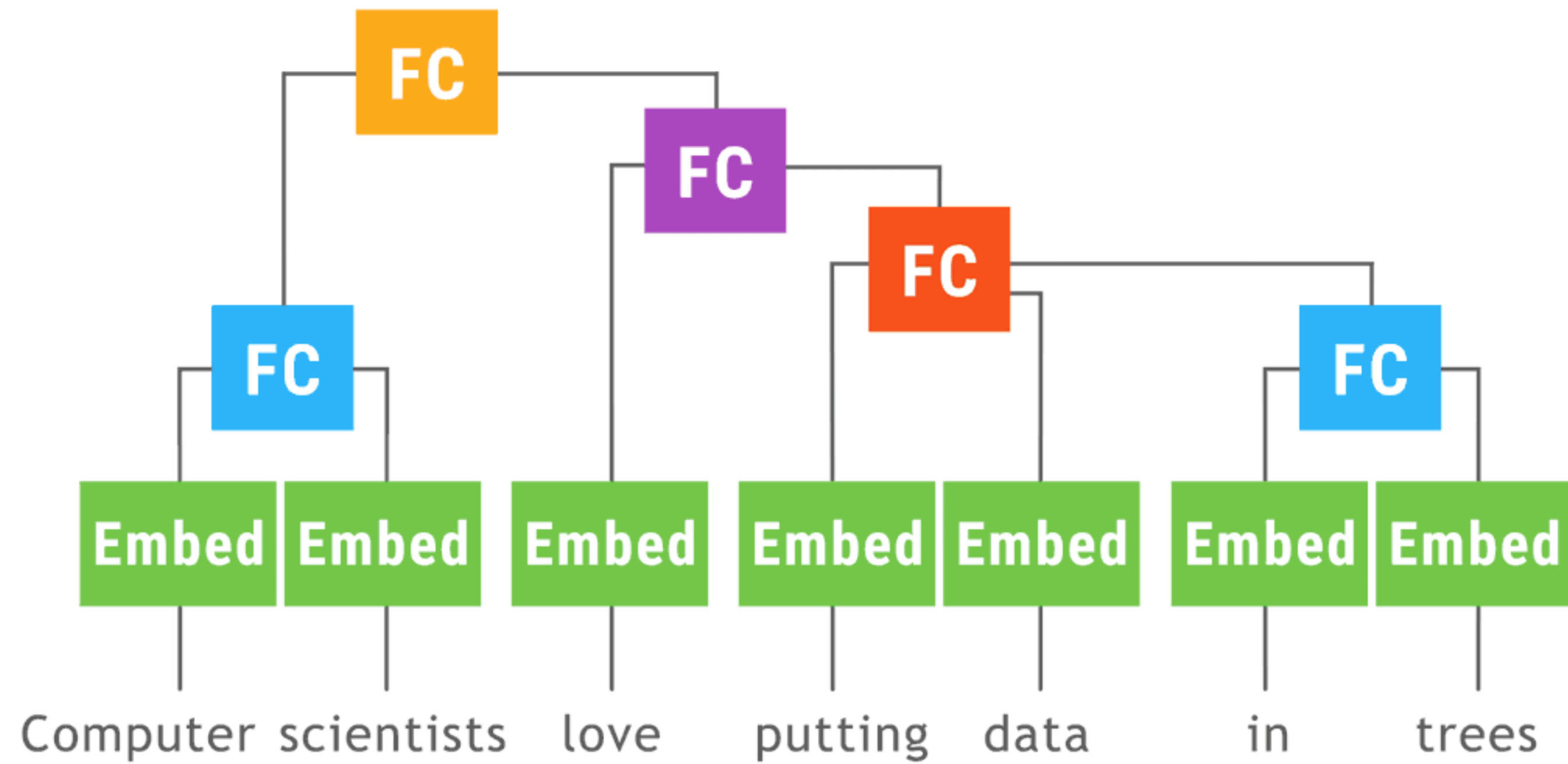
◆ Cons:

- ❖ Doesn't support dynamic programs

- ◆ EC2 P3.8xlarge: 4 Tesla Volta
- ◆ Resnet 50 training



Folding

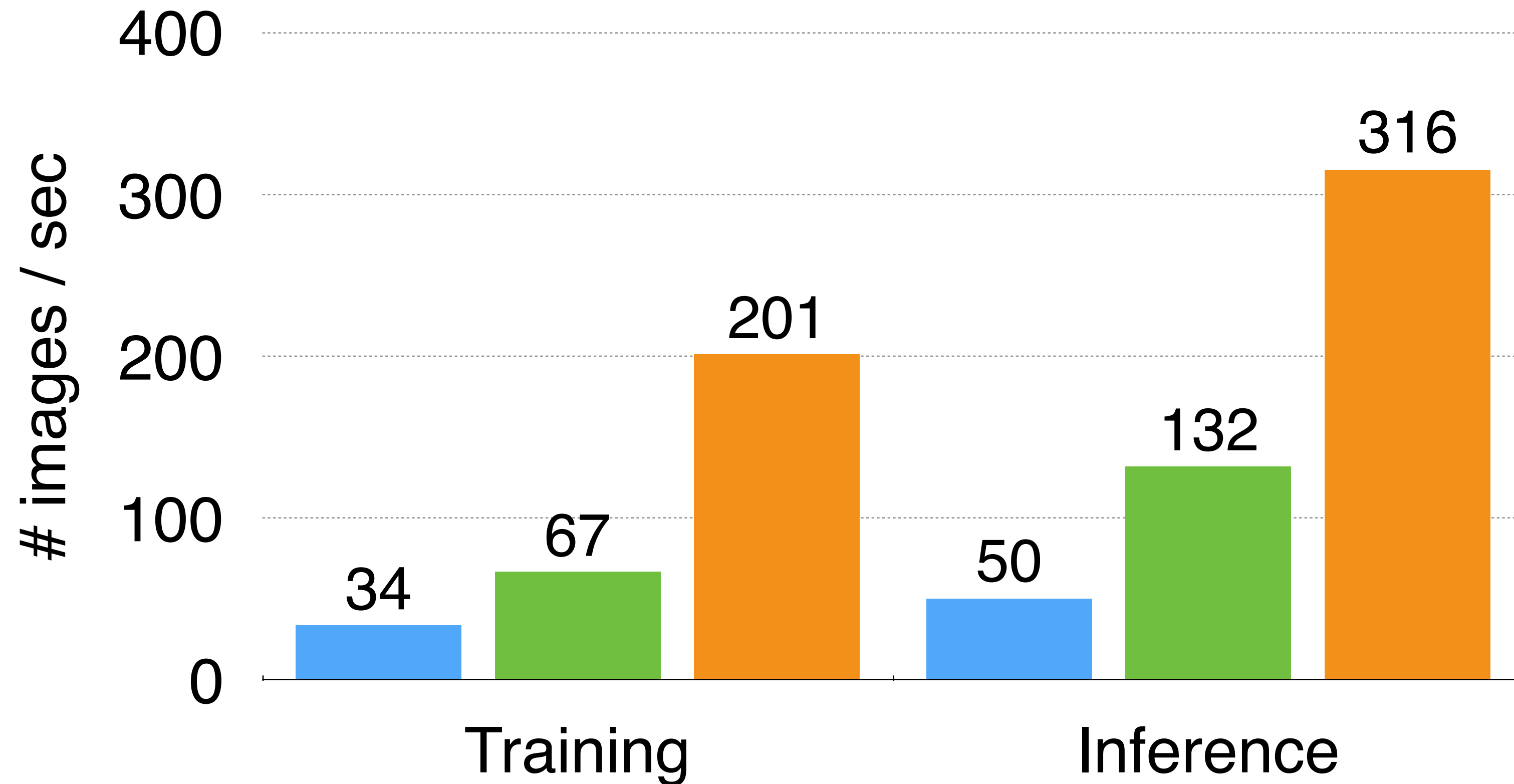


[Figure adapted from Moshe Looks]

[Looks, et.al 2017]

Folding Performance

■ w/o fold ■ w/o fold, hybridized ■ w/ fold



- ◆ EC2 C4.8xlarge
- ◆ Tree LSTM

GPUs

Summary

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