第8章
Word2vec

14T4027Y 熊谷佳奈

chainerによるword2vecの実装

- インポート

w2v.py

```
>>> import numpy as np
>>> import chainer
>>> from chainer import cuda, Function, \
...                 Variable, optimizers, serializers, utils
>>> from chainer import Link, Chain, ChainList
>>> import chainer.functions as F
>>> import chainer.links as L
>>> import collections  # (注意) これが必要です
```

chainerによるword2vecの実装

- コーパス中の単語にidをつける

```
>>> index2word = {}
>>> word2index = {}
>>> counts = collections.Counter()
>>> dataset = []
>>> with open('ptb.train.txt') as f:
...     for line in f:
...         for word in line.split():
...             if word not in word2index:
...                 ind = len(word2index)
...                 word2index[word] = ind
...                 index2word[ind] = word
...             counts[word2index[word]] += 1
...             dataset.append(word2index[word])
>>> n_vocab = len(word2index)
>>> datasize = len(dataset)
```

chainerによるword2vecの実装

- サンプル生成器の作成

ノイズ分布

$$p(w) = \frac{U(w)^{0.75}}{\sum_{v=1}^{V} U(w_v)^{0.75}}$$

```
>>> from chainer.utils import walker_alias
>>> cs = [counts[w] for w in range(len(counts))]
>>> power = np.float32(0.75)
>>> p = np.array(cs, power.dtype)
>>> sampler = walker_alias.WalkerAlias(p)
```

chainerによるword2vecの実装

- Sampler.sample()

```
>>> sampler.sample(5)
array([  47,  737, 5022,  219,   25], dtype=int32)
```

chainerによるword2vecの実装

- MyW2V

```
>>>class MyW2V(chainer.Chain):
...     def__init__(self, n_vocab, nunits):
...           super(MyW2V, self).__init__(
...                 embed=L.EmbedID(n_vocab, nunits),
...           )
...     def__call__(self, xb, yb, tb):
...           xc = Variable(np.array(xb, dtype=np.int32))
...           yc = Variable(np.array(yb, dtype=np.int32))
...           tc = Variable(np.array(tb, dtype=np.int32))
...           fv = self.fwd(xc,yc)
...           return F.sigmoid_cross_entropu(fv, tc)
...     def fwd(self, x, y):
...           xv = self.embed(x)
...           yv = self.embed(y)
...           return F.sum(xv * yv, axis=1)
```

chainerによるword2vecの実装

- モデルと最適化アルゴリズムの設定

```
>>> demb = 100
>>> model = MyW2V(n_vocab, demb)
>>> optimizer = optimizers.Adam()
>>> optimizer.setup(model)
```

- 単語ペアの集合の作成

```
>>> ws = 3          # window size
>>> ngs = 5         # negative sample size
```

```
>>> def mkbatset(dataset, ids):
...     xb, yb, tb = [], [], []
...     for pos in ids:
...         xid = dataset[pos]
...         for i in range(1,ws):
...             p = pos - i
...             if p >= 0:
...                 xb.append(xid)
...                 yid = dataset[p]
...                 yb.append(yid)
...                 tb.append(1)
...                 for nid in sampler.sample(ngs):
...                     xb.append(yid)
...                     yb.append(nid)
...                     tb.append(0)
...             p = pos + i
...             if p < datasize:
...                 xb.append(xid)
...                 yid = dataset[p]
...                 yb.append(yid)
...                 tb.append(1)
...                 for nid in sampler.sample(ngs):
...                     xb.append(yid)
...                     yb.append(nid)
...                     tb.append(0)
...     return [xb, yb, tb]
```

chainerによるword2vecの実装

- パラメーターの更新

```
>>> bs = 100       # batch size
>>> for epoch in range(10):
...     print('epoch: {0}'.format(epoch))
...     indexes = np.random.permutation(datasize)
...     for pos in range(0, datasize, bs):
...         print epoch, pos
...         ids = indexes[pos:\
...           (pos+bs) if (pos+bs) < datasize else datasize]
...         xb, yb, tb = mkbatset(dataset, ids)
...         model.cleargrads()
...         loss = model(xb, yb, tb)
...         loss.backward()
...         optimizer.update()
```

chainerによるword2vecの実装

- 保存

```
>>> with open('myw2v.model', 'w') as f:
...     f.write('%d %d\n' % (len(index2word), 100))
...     w = model.embed.W.data
...     for i in range(w.shape[0]):
...         v = ' '.join(['%f' % v for v in w[i]])
...         f.write('%s %s\n' % (index2word[i], v))
```

chainerによるword2vecの実装

- 例