A TensorFlow Tutorial: Sentiment Classification using LSTMs

Kunal Kotian

June 28, 2018

Key aspects covered by the tutorial

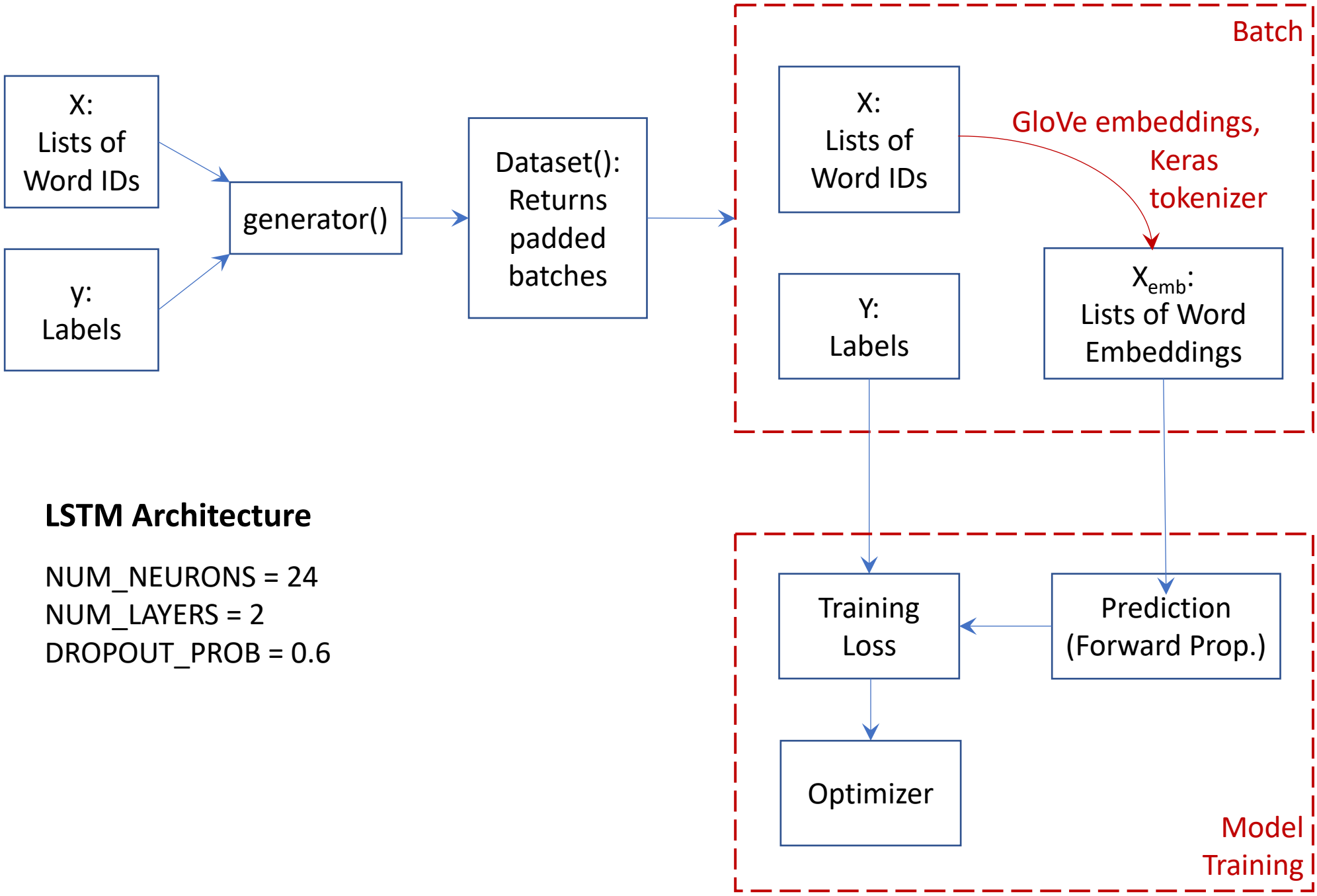
- Minimal text preprocessing
  - Contraction replacement + tokenization
- Using word embeddings
- TensorFlow's Dataset API for batching and loading data
- Efficient handling of variable length input
- Example used
  - Yelp dataset of restaurant reviews
    - Polarized reviews selected (1 or 2 stars = negative; 5 stars = positive)
    - 400 to 500 characters per review

Using Word Embeddings in tf

- Tokenizer converts text strings to lists of integer tokens
  - Maps word <---> integer token
  - Integer tokens = ids with no special significance

```
Text review:
 Very disappointed. My husband enjoys fine food and we eat out often. But after telling staff he did not enjoy his me
al, they we are rude and argumentative. We we are treated like we we are trying to pull a con and get some money off
our bill, when my husband had actually found the meat to be tough and just about inedible. Will never go back and thi
nk this company need to retrain their staff to show customers a bit of respect and remember where their incomes comes
from.

Integer sequence representation:
 [35, 183, 17, 267, 3988, 565, 16, 2, 8, 105, 43, 559, 21, 93, 1512, 90, 107, 55, 10, 343, 261, 135, 19, 8, 12, 331,
2, 8, 8, 12, 1101, 42, 8, 8, 12, 404, 5, 1941, 4, 2427, 2, 51, 85, 280, 141, 37, 500, 54, 17, 267, 24, 348, 288, 1, 1
81, 5, 30, 972, 2, 40, 67, 1420, 34, 77, 48, 41, 2, 177, 15, 1206, 307, 5, 49, 90, 5, 699, 373, 4, 274, 11, 2548, 2,
675, 248, 49, 463, 53]
```



LSTM Architecture

NUM_NEURONS = 24
NUM_LAYERS = 2
DROPOUT_PROB = 0.6

Many-to-one LSTM for Sentiment Classification

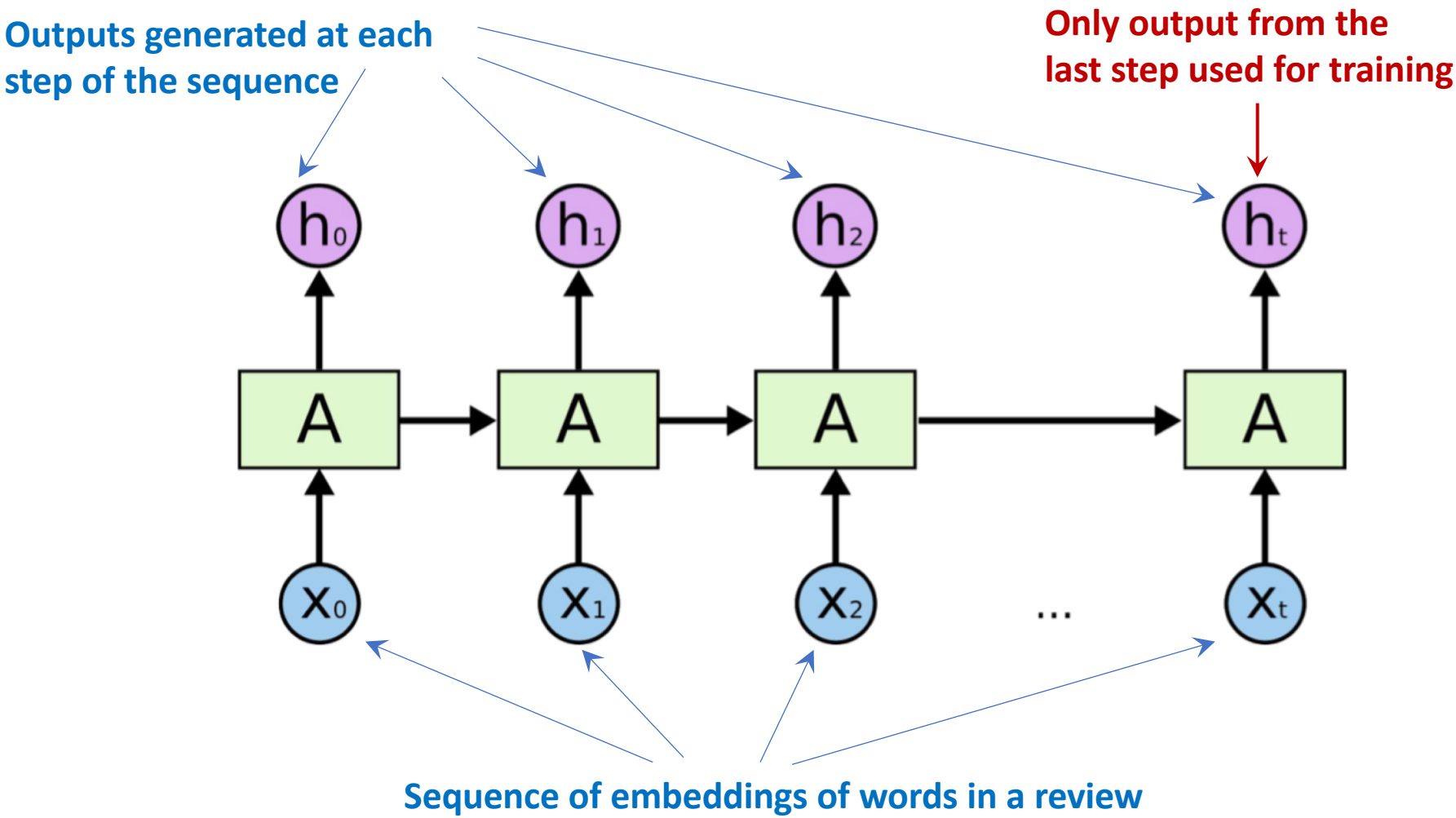


Image source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Variable Length Input Considerations

- Reviews have an unequal number of words
- Zero padding at **batch level**
  - Sequence length set by the longest review in each batch
- Identify **last 'relevant' output** for each review
  - For loss calculation and sentiment prediction

Only LSTM output at this step is used

I	**like**	**trains.**	**0**	**0**
No	**really,**	**I**	**love**	**trains.**
Trains	**are**	**fun**	**right?**	**0**