Natural Language Processing

Today's agenda:

- Review
  - Preprocessing
  - Word embeddings
- Sequence to sequence
  - Encoder/Decoder
  - Neural Machine Translation (NMT)
- Tutorial (RNN and Seq2seq)

Natural Language Processing (NLP)

NLP is yet another area that Deep Neural Networks excel at and includes such tasks as:

- Classification
  - Topic identification
  - Sentiment analysis, ...
- Sequence to sequence
  - Language modeling
  - Machine translation
  - Summarization, ...

Preparing text data

Raw text is not something we can feed to a Neural Network. We thus have to preprocess text data to efficiently encode the same information numerically in the form of vectors.

- Standardization
- Tokenization
- Indexing
- Embedding

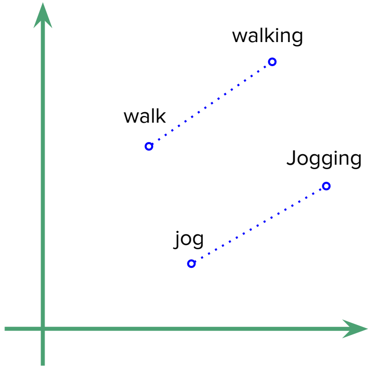
Preparing text data

In Keras we would use the TextVectorization layer:

- Convert to lowercase and remove punctuations
- Tokenize by words/custom tokenization
- Index via the adapt() method

Word embeddings

With a little bit of feature engineering we can drastically reduce the dimension of the “word space” such that the relative location of the words encodes their relationship:



Embedding layer

The Embedding layer in Keras randomly assigns “dense” vectors to our indexed words. These vectors are trainable parameters that learn the relationships among words during training.

Pretrained embeddings: Word2Vec

> “You shall know a word by the company it keeps.”
>
> *J.R. Firth*

Pretrained embeddings such as Word2Vec encode words by their frequent companions.

- Encode synonyms and word relationships
- Capture context to a limited degree
- Reduce the dimension of embedding space

Sequence to sequence (Seq2seq)

Prior to 2014, the state of the art in Machine Translation was highly engineered Statistical Machine Translation models.

- 2014: The first Seq2seq translation model (NMT) appears
- 2016: Google translate switches to NMT

Sequence to sequence (Seq2seq)

Neural Machine Translation models have a few advantages over their predecessors:

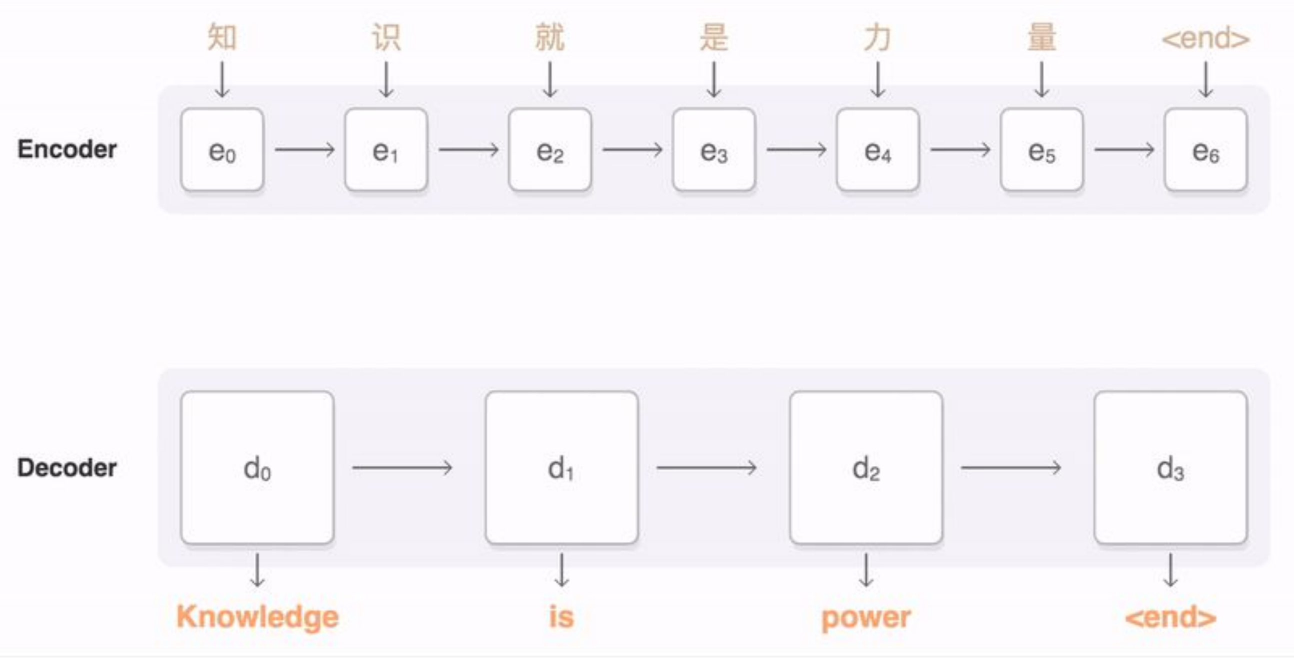
- Better performance (fluency, context, phrase similarity)
- End-to-end deep neural net
- Minimal engineering involved

Sequence to sequence (Seq2seq)

Some shortcomings of NMTs:

- Out of vocabulary (OOV) words
- Maintaining context over longer texts
- Sentence meaning accuracy
- Underrepresented languages
- Inherent biases such as gender bias

Encoder-decoder structure



From google.github.io/seq2seq

Encoder

The purpose of the encoder is to extract relevant features and pass them on in sequential form. The hope is that the extracted features encode the context and meaning of the text in a language-independent fashion.

The encoder is usually a stacked RNN with two gated recurrent layers.

Decoder

The purpose of the decoder is to describe the extracted features in a different language.

- Greedy decoding
- Beam search decoding

Questions?