Neural Machine Translation

Sargur N. Srihari
srihari@cedar.buffalo.edu

This is part of lecture slides on Deep Learning:
http://www.cedar.buffalo.edu/~srihari/CSE676

Topics in NLP

1. N-gram Models
2. Neural Language Models
3. High-Dimensional Outputs
4. Combining Neural Language Models with n-grams
5. Neural Machine Translation
6. Historical Perspective

Topics in Neural Machine Translation

- Overview
- Using an Attention Mechanism and Aligning Pieces of Data

The machine translation task

- It is the task of reading a sentence in one natural language and emitting a sentence with an equivalent meaning in another language
- At a high level, there is a component that proposes many candidate translations
  - Many translations will not be grammatical

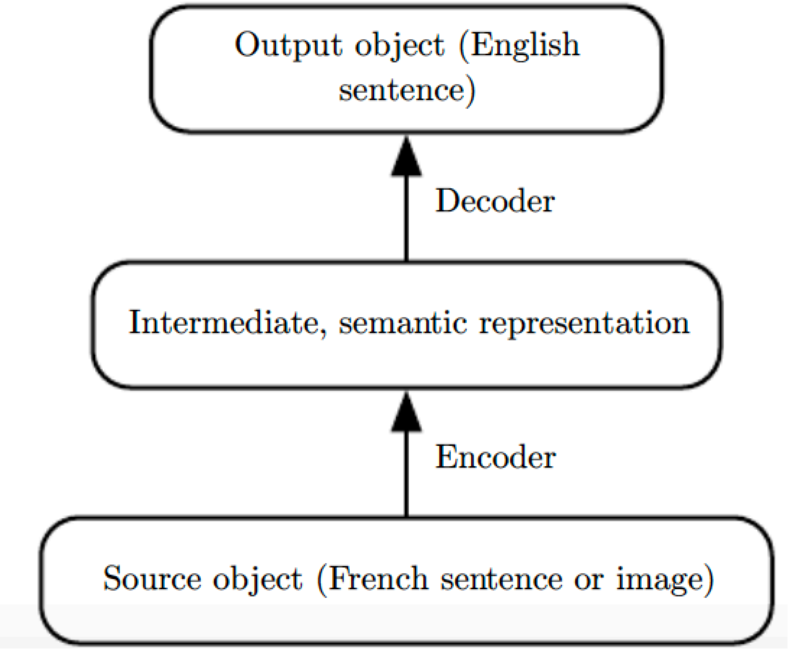
History of Machine Translation (MT)

- Early MT systems used *n*-gram models
  - Including maximum entropy language models
  - Report probability of a natural language sentence
- An MLP MT produces a sentence given input
  - Produces a conditional distribution given context C
    - Where C is a single variable or a list of variables
  - An MLP scores a phrase $\mathrm{t}_1,..,\mathrm{t}_\mathrm{k}$ given a phrase $\mathrm{s}_1,..,\mathrm{s}_\mathrm{n}$ by estimating $P(\mathrm{t}_1,..,\mathrm{t}_\mathrm{k}\ |\mathrm{s}_1,..,\mathrm{s}_\mathrm{n}\ )$

MLP versus RNN

- MLP requires inputs to be preprocessed to be of fixed length
- RNN provides ability to accommodate variable length inputs and variable length outputs
- Model first reads an input sequence and emits a data structure that summarizes the input sequence
  - We call this summary the "context" C
- An RNN then reads context C and generates a sentence in the target language

The encoder-decoder architecture


Output object (English sentence)
Decoder
Intermediate, semantic representation
Encoder
Source object (French sentence or image)

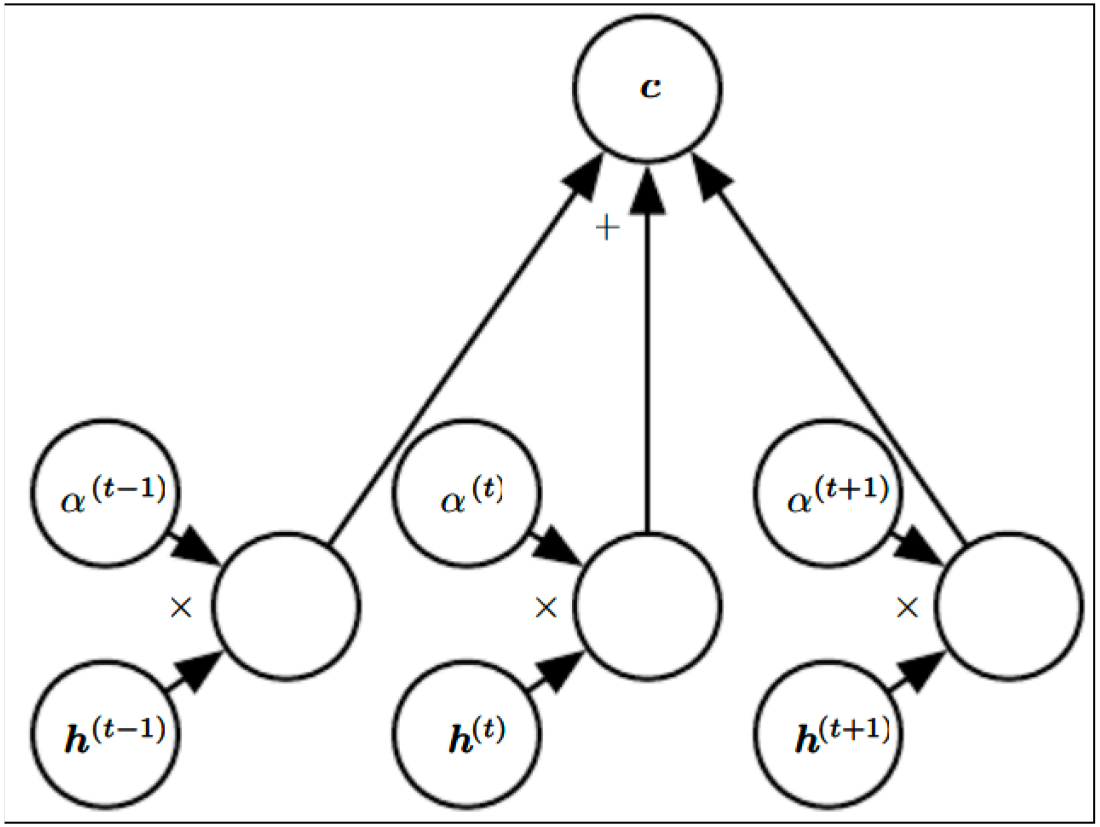
Map back and forth between a surface representation (sequence of words) and a semantic representation

By using the output of an encoder of data from one modality (encoder mapping French sentences to hidden representations capturing the meaning of sentences as input to a decoder for another modality (such as the decoder mapping from hidden representations capturing the meaning of sentences to English)

Using an attention mechanism and aligning pieces of data

- Using a fixed-size representation to capture all the semantic details of a very long sentence of 60 words is very difficult
- While it can be done by an RNN trained well-enough and long enough
- More efficient approach is to read the whole sentence or paragraph (to get gist or context) then produce translated words one at a time each time focusing on a different part of the input sentence

A modern attention mechanism



It is essentially a weighted average.

A context vector $\boldsymbol{c}$ is formed by taking a weighted average of feature vectors $\boldsymbol{h}^{(t)}$ and weights $\alpha^{(t)}$

Weights $\alpha^{(t)}$ are produced by the model itself

They are usually values in the interval $[0,1]$ and are intended to concentrate around one $\boldsymbol{h}^{(\mathrm{t})}$ so that the weighted average approximates reading that one specific time precisely

Weights $\alpha^{(t)}$ are produced by applying a softmax function to the relevant scores emitted by another portion of the model

Cost of attention mechanism

- It is more expensive computationally than directly indexing the desired $\boldsymbol{h}^{(t)}$
- But direct indexing cannot be trained with gradient descent
- The attention mechanism based on weighted averages is a smooth, differentiable approximation that can be trained with existing approximation algorithms

Components of attention-based system

- An attention-based system has 3 components:
  1. A process that reads raw data (such as source words in a source sentence) and converts them into distributed representations with one feature vector associated with each word position
  2. A list of feature vectors storing the output of the reader. This can be thought of as memory containing a sequence of facts, which can be retrieved, not necessarily in order
  3. A process that exploits the content of the memory to sequentially perform a task at each time step having the ability to put attention on one memory element
- The third component generates the translated sentence

Relating word embeddings

- When words written in one language are aligned with corresponding words in a translated sentence in another language, we can relate corresponding word embeddings
- Earlier work:
  - Learn translation matrix relating word embeddings in a language with embeddings in another
    - Yielding lower alignment error rates than traditional methods based on frequency counts in phrase tables
  - Cross-lingual word vectors
    - Extension: more efficient cross-lingual alignment allows training on larger datasets