Encoder-decoder paradigm for RNNs: with applications in neural machine translation, speech recognition, and image captioning

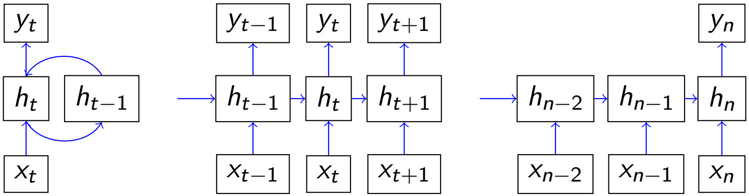
Arvid Frydenlund

October 5, 2015

Overview

- RNNs and RNN language models
- Alignment problem in translation
- Encoder-decoder paradigm (also called sequence to sequence)
- Applications
- My work

RNNs



a) RNN at t *b) unrolled RNN in time* *c) single output RNN*

RNNs

- Allow for arbitrary context length
- Neural language models are a classifier over $|V|$ classes (generally between 20,000 - 600,000).
- Recurrently read in words, predict word at the end
  - (the, cat, sat, on, the) $\rightarrow$ (mat)
  - $\operatorname{argmax}_{w_n} P(w_n | w_1, .., w_{n-1})$
- RNNs can also make a prediction after every input
  - (frame 1, frame 2, frame 3, frame 4) $\rightarrow$ (k,k,a,a)
- If n sized input, an RNN can have m outputs where $m \leq n$

Alignment problem

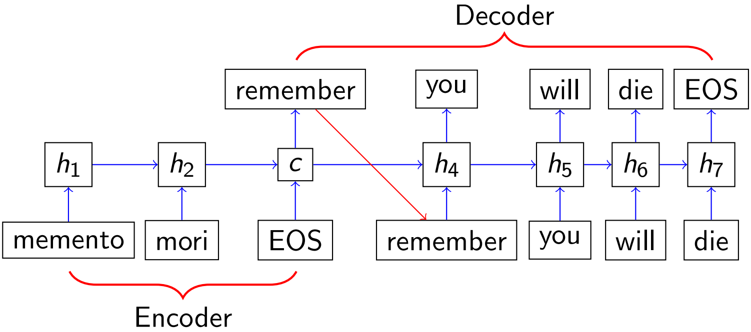
- What if we want $m > n$?
- If m is dynamically between 1 and n, how do we determine when to output?
- Machine translation requires an arbitrary alignment
- Need a generalized function $(x_1, ..., x_n) \rightarrow (y_1, ..., y_m)$ for arbitrary n and m

Solution: Encoder-decoder paradigm

- Encoder reads in input into a fixed length representation, c
- Decoder generates arbitrary length output conditioned on c
- Each decoder output at t serves as input to the decoder at $t+1$
- Loops until $< \mathrm{EOS} >$ symbol is generated
- $P(y_1, .., y_m | x_1, .., x_n) = \prod_{t=1}^{m} P(y_t | c, y_1, .., y_{t-1})$
- Encoder and decoder are trained in an end to end fashion
- Not limited to RNNs (CNN encoders are common)

- *Sequence to sequence learning with neural networks*, Sutskever *et al.*, NIPS, 2014
- *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, Cho et al., EMNLP, 2014
- *Two recurrent continuous translation models*, Kalchbrenner and Blunsom, ACL, 2013

Neural machine translation encoder-decoder



Neural machine translation (NMT)

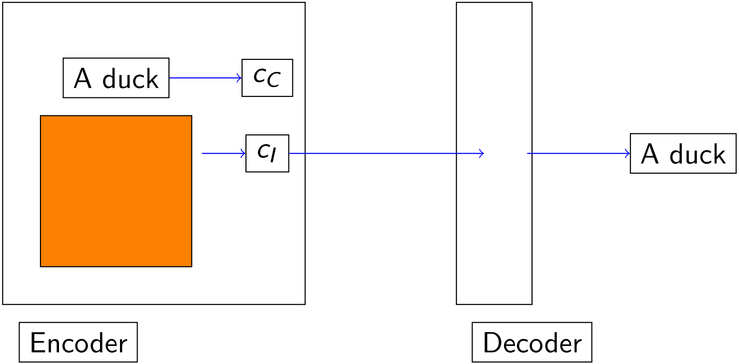
- Encoder creates a sentence embedding, c, in one language and the decoder is a language model in another language which is conditioned on c
- $P(f|e) = \prod_{t=1}^{m} P(f_t|f_1, ..., f_{t-1}, c)$, a direct modelling of english to french instead of $P(e|f)P(f)$
- NMT getting state-of-the-art results in only 2 years (using one major extension called attention mechanisms)
- Generally use stacked bi-directional RNNs, 8 layers deep $(4 + 4)$, and projection layers, LSTM or GRU units
- Small vocabulary $20,000 - 30,000$ words for both the input and output languages
- *Montreal Neural Machine Translation Systems for WMT15*, Jean *et al.*, Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015

Other applications

- Sentence summarization
  - Compress paragraph or article into a fixed length representation, then decoder generates a sentence that summarizes it
  - Does *abstractive* summarization instead of *extractive*
  - *A neural Attention Model for Abstractive Sentence Summarization*, Rush et al., preprint, 2015
- End-to-end RNN speech recognition
  - Compress a speech utterance into a fixed length representation, then decoder generates a sequence of phones or letters
  - Very hard to compress a whole utterance: main challenge is designing a good encoder

Other applications - Image captioning

- Compress a caption to c_C and try to regenerate the caption
- Compress image to c_I, such that $c_C \approx c_I$



- *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, Xu *et al.* (Montreal and UofT people), ICML, 2015

My research - morphological language models

- Closed vocabularies: a set of acceptable input words and a set of output words (classes)
- Causes out of vocabulary errors, requires OOV token
- Compositional models: don't assume words as a base unit
- Helps solve open input vocabulary problem
- Helps data sparsity: *re-input*, *re-input-ed*, *input-ed*
- Does it help the open output vocabulary problem?
- (I, have, been, work, -ing, on) $\rightarrow$ (re, -input-, ing)
- Reduces the number of classes: *cat*, *cat-s*
- Why: Open output vocabularies not really done so far
- Why: Morphological language models have not been used for speech recognition