The Goldilocks Principle

Abstract

* They distinguish between syntactic function words (like prepositions) and words that carry more semantic meaning (like nouns)
* Models which can store explicit representations of long-term contexts out perform state of the art models for predicting semantic words, not syntactic words
* The amount of text encoded in a single memory representation is highly influential to the performance
* You need to encode a sweet spot between single words and full sentences (Goldilocks Principle)

Introduction

* Based off a dataset “Children’s Book Test” AKA “CBT” which makes predictions about different types of missing words in children’s books given nearby sentences as context
* It has 20 buildup sentences and then…
  + Ex: the \_\_\_ jumped over the moon
    - choices: {fox, hen, …., cow, pig} (10 choices
    - answer: cow
* humans taking the test rely on a wider context to make predictions about named entities (nouns), more than just the question (query) itself
* LSTMs are really good at predicting syntactic words
* Memory networks are better at predicting nouns

CBT

* Developed to see how language models can explore a wider linguistic context
* 20 buildup sentences and then a query
* Because traditional language models predict the next word in a sequence, they naturally put more emphasis on more frequently used words

Studying Memory Representation w/ End to End Memory Networks

* Briefly explained in paper
* Maps sequences of words to one-hot representations the size of the vocabulary (word2vec may be an improvement)
* Lexical memory method – each word occupies a separate slot in the memory
* Window memory – each phrase corresponds to a window of text from the context, the memory slots are filled w/ windows of words
* Sentential memory – phrases of memory networks correspond to complete sentences of the context (dimensionality reduction method)
* Sentential memory was the original implementation of end to end memory networks when the were introduced for NLP
* We embed our queries and memories into the same dimensionality
* We feed the embedding space through a SoftMax layer which acts as an attention mechanism
* This process is done for all the memories stored (20 in this paper)
* Achieved subhuman performance, human performance was around 80%, the networks here achieved 66-70% (except for LSTMs when it came to predicting prepositions and verbs)

My takeaways

* Encoding variables on the scale of phrases as opposed to words may be the next way forward, a complete idea can be represented better in a phrase than in a word
* If the goal of these networks to predict language would be to organically string together ideas into a narrative, it may make sense to encode the phrases and not the words
  + May also help it train faster too