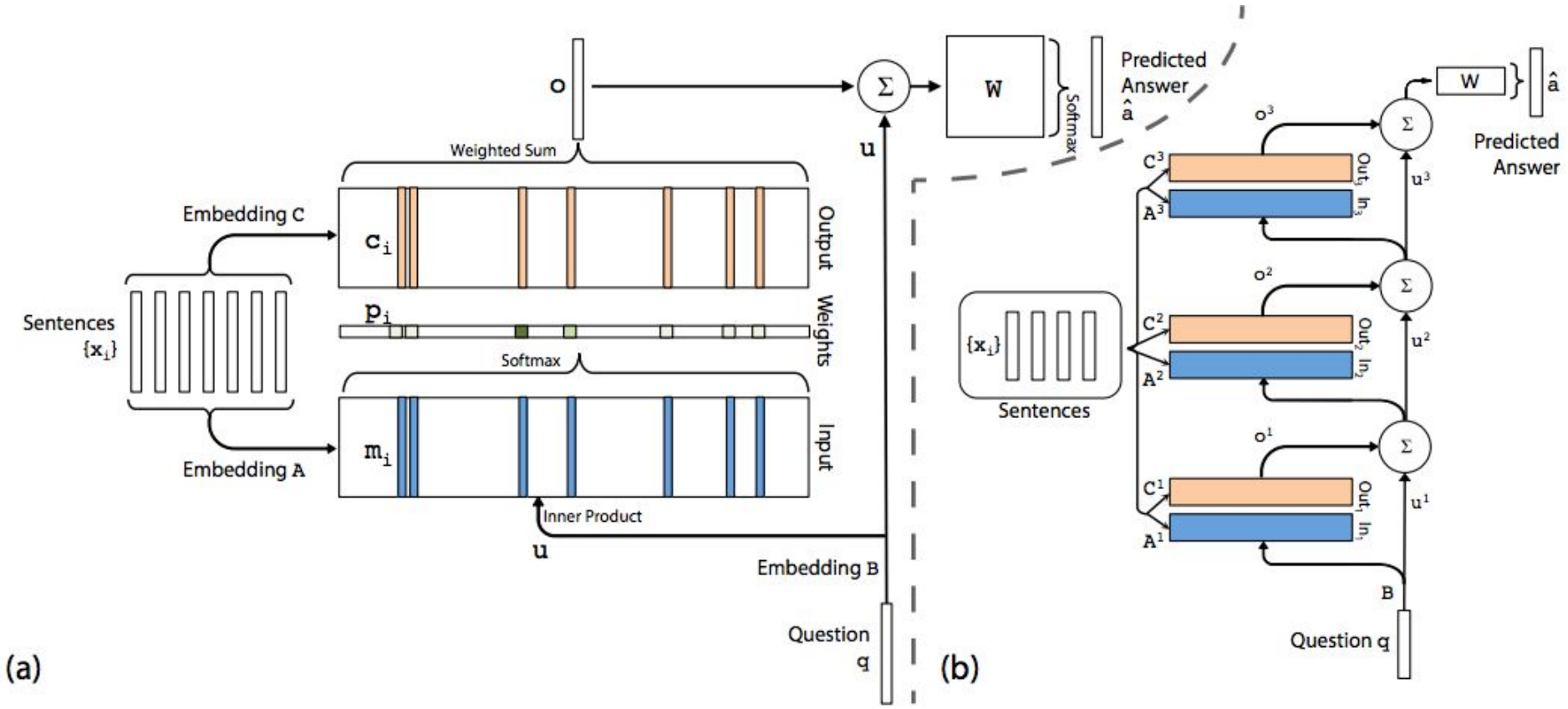
End to end memory networks as language models

Implementation and Investigation

William Jack

End to end memory network as presented by Sukhbaatar et al.



- Originally designed for question answering tasks.
- Performance dramatically increased by making recurrent “hop” operations.

One hop memory network

- Took 30 words as context, goal of predicting the next word, used 512kb chunks of the text8 corpus for training and testing
- Trained for 2 epochs using Adagrad
- Train perplexity of **369.54**, test perplexity of **1870.63**
- Example generated sentences from the test set (first generated by picking max. probability next word, second generated by drawing from the probability distribution over all words):
  - "ribbon worms the sipuncula and several phyla that have a fan of cilia around the mouth called a lophophore these were traditionally grouped together as the lophophorates but it now | **being found in the united states and the most of the united states and the most"**
  - "mankind s place in the universe hinges on the notion that all h or god is the only true reality there is nothing permanent other than him god is considered | **during non famous unsolved also immaterial csm numeric the great dihydrogen skeleton the king mentioned resulted"**

Three hop memory network -- SGD

- Used stochastic gradient descent with a learning rate of .01
- Train perplexity of **4797.85**, test perplexity of **5776.39** after 2.5 epochs (~2.5 days)
- Example generated sentences from the test set (first generated by picking max. probability next word, second generated by drawing from the probability distribution over all words):
  - zero mi a great rift valley also extends along the ridge over most of its length the depth of water over the ridge is less than two seven zero zero | **in in in in in in in in in in in in in in in in in**
  - six metres one two eight eight one ft the greatest depth eight six zero five metres two eight two three two ft is in the puerto rico trench the width | **rivers that a zero blood ruling in swan occupations part cataclysmic according a albedo in ism**

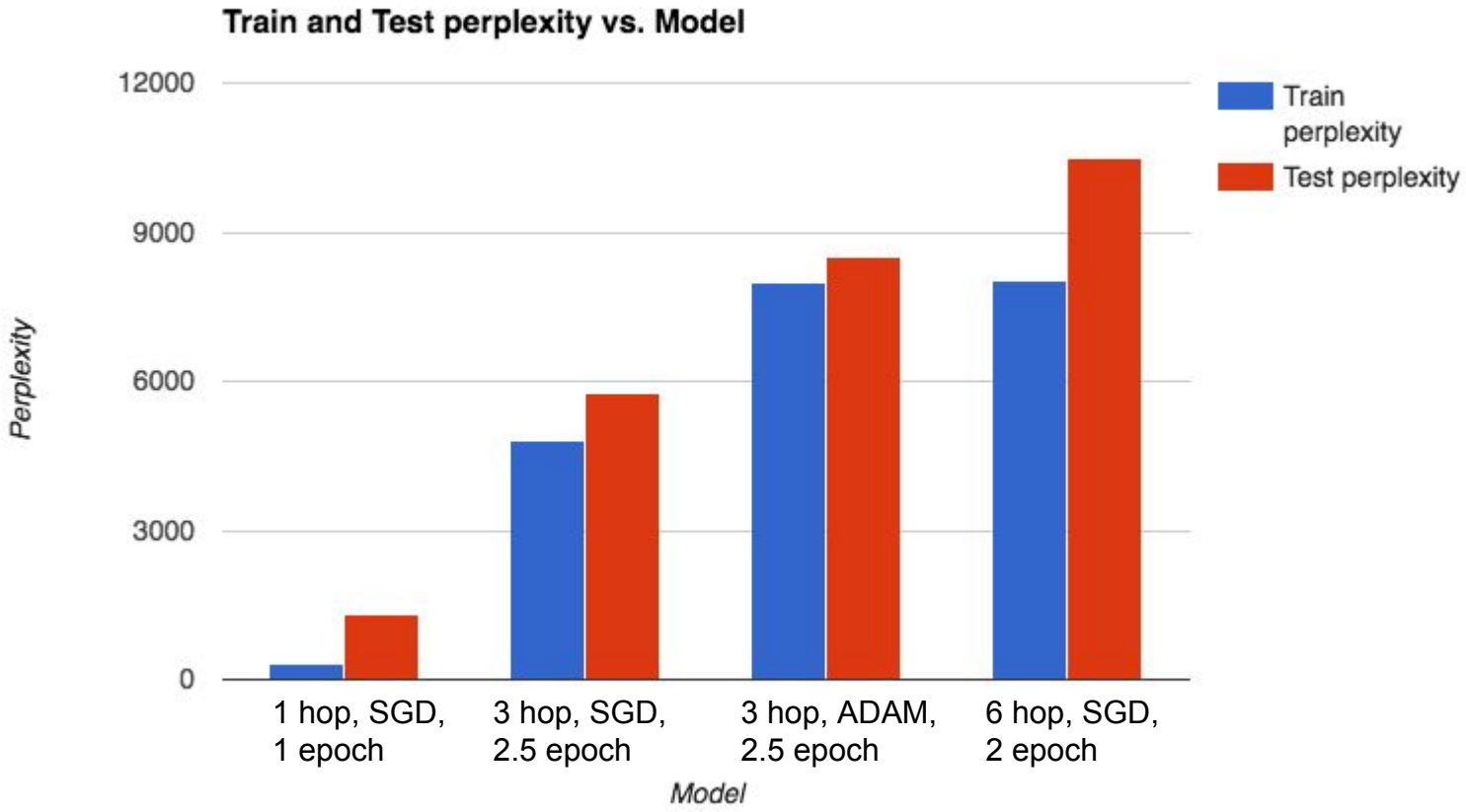
Three hop memory network -- ADAM

- Used ADAM optimizer with a learning rate of .01, beta1=0.9, beta2=0.999, epsilon=1e-08
  - 'appropriate for problems with very noisy or sparse gradients' (Kingma, Ba 2015)
- Train perplexity of **7969.99**, test perplexity of **8519.64** after 2.5 epochs (~2.5 days)
- Example generated sentences from the test set (first generated by picking max. probability next word, second generated by drawing from the probability distribution over all words):
  - premise asserts if and only if rather than if similarly the converse of a statement can be validly assumed to be true so long as the if and only if | **be such in the the three and the the three three three three three three three**
  - but i can tell you what they will use in the fourth rocks einstein was a supporter of zionism he supported jewish settlement of the ancient seat of judaism and | **stimulatory implementation incomes usgs grandibracteata mercy specialties for zero frogs anti zero that three five anglophone**

Six hop memory network implementation

- Used stochastic gradient descent with learning rate of .01
- Trained for 2 epochs (appox. 2.5 days)
- Train perplexity of **8009.86** test perplexity of **10508.944**
- Example generated sentences from the test set (first generated by picking max. probability next word, second generated by drawing from the probability distribution over all words):
  - "s opposite other examples of modest proposals modest proposals and other literary hoaxes report from iron mountain sokal affair miscegenation origin of the word dihydrogen monoxide jack thompson attorney has | **a a a a a a a a a a a a a a a a a a**"
  - "continent shaking wars he did indeed maintain his aloof position of minding not the times but the eternities schopenhauer on women schopenhauer is also famous for his essay on women | **repeatedly mammary widow milligray ships bernazzani littoral reign saw buzkashi equalization moves caribou layer filament snipe**"

Comparison between models



- More hops leads to significantly longer training time.
- The difference in train and test perplexity for more-hop models is small, so there is much hope for improvement with more training.
- SGD learns significantly faster than ADAM.

Sukhbaatar et al. Results

	# of hidden	# of hops	Text8 memory size	Valid. perp.	Test perp.
RNN	500	-	-	-	184
LSTM	500	-	-	122	154
SCRN	500	-	-	-	161
MemN2N	500	2	100	152	187
	500	3	100	142	178
	500	4	100	129	162
	500	5	100	123	154
	500	6	100	124	155
	500	7	100	118	**147**
	500	6	25	131	163
	500	6	50	132	166
	500	6	75	126	158
	500	6	100	124	155
	500	6	125	125	157
	500	6	150	123	154

Taken from Sukhbaatar's 'End-to-End Memory Networks'

- Used a dataset ~20x bigger.
- Trained for ~50 epochs on this dataset rather than ~2
- Used 100 words as context rather than 30,
- Used embedding dimension 500 rather than 300.
- Compromises were made in our models for increased speed due to time and cost restraints.

Character Level Language Model

- SGD, lr = .01, 1 epoch, minibatch size = 1 to help avoid local minima.
- Gradient normalization to < 15
- 12 memory hops, 400 characters of context, 100 embedding dimensions.
- Train perplexity: **21.40**, Test perplexity: **18.095**
- ‘... and assakenois or assaceni assacani asscenus the aspasios assakenois ashvakas cavalrymen is stated to be another name for the kambojas because| **z spyx loa wh mynmteieqe badm liccoeoonrvnoeaglj ify if lusjearehh q xfeew idnbtol ltegrisniaikwc aa**’
- Max prob generation predicts “ ” for everything.
- Has not learned a lot, though the gradients were nonzero and the cross entropy was decreasing throughout training.

Next Steps

- Run models for more epochs, do so on GPUs that support tensorflow.
- Try feeding word vectors produced from word2vec into the model to decrease the size of the model, and have it learn less about how to embed words, and more how to relate them.
- Implement learning rate annealing every 25 epochs as described in Sukhbaatar's paper.

Sources cited:

"End-To-End Memory Networks" Sukhbaatar et al. 2015. arXiv:1503.08895