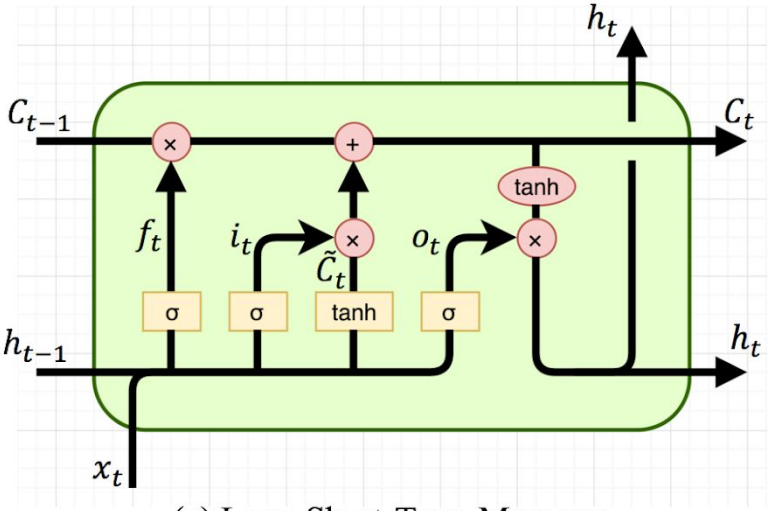
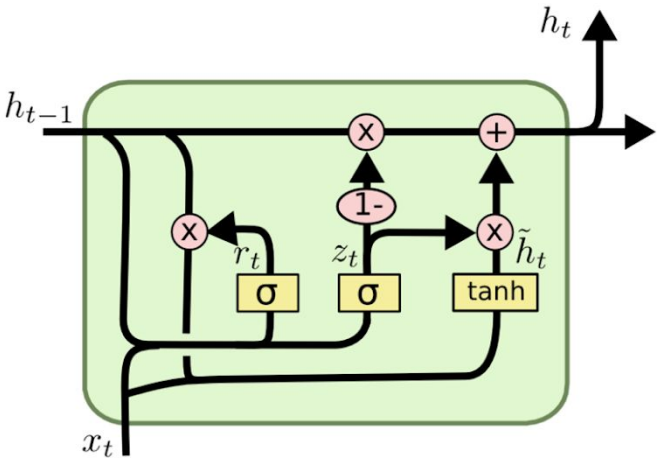
Language Modeling with RNNs

Language Modeling

LSTM and GRU



(a) Long Short-Term Memory



(b) Gated Recurrent Unit

LSTM: A Search Space Odyssey, Klaus Greff et al. (IEEE Tran. on N.N. and Learning Systems 2016)

Capacity and Trainability in Recurrent Neural Networks, Jasmine Collins et al. (ICLR 2017)

The Unreasonable Effectiveness of Recurrent Neural Networks (Karpathy 2015)

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS)[http://www.humah.yahoo.com/guardian.

Tips & Tricks (Melis et al. (ICLR 2018); Merity et al. (ICLR 2018))

- Weight Decay (~1 / (2 * N))
- Weight Tying (Embeddings and decoder)
- Word Dropout ([0, 0.2])
- Dropout (embeddings, layers, output) ([0, 0.9])
- Weight Dropout ([0, 0.9])
- Variational Dropout (Gal & Ghahramani NIPS 2016)
- Independent embedding size and hidden size
- ...

Tokenization & Sub Words

Word level vs character level

- Word level (e.g. NLTK, OpenNTP, fairseq, etc...) => perplexity
- Character level => byte / character

Sub Words:

- BPE, Sennrich et .al (ACL 2016) BPE on Wikipedia (C++: Lample FastBPE)
- WordPiece, Schuster & Nakajima (ICASSP 2012) (PyTorch: HuggingFace BertTokenizer)
- Subword Regularization, Kudo (ACL 2018) (Python: Google SentencePiece)

Decoding with LSTMs

How to generate samples with highest likelihood ?

- Greedy: fast and easy but inefficient
- Top K sampling: stochasticity
- Beam Search (PyTorch: Facebook fairseq): heavier (time & memory), better samples, no stochasticity
- Stochastic Beams and Where to Find Them (Kool et al. ICML 2019) (?)