Recurrent Neural Network for Language Modeling

EE596/LING580 -- Conversational Artificial Intelligence

Hao Cheng
University of Washington

Language Model Basics

- $P(\boldsymbol{w}) = \prod_{i=1}^{T} P\left(w_j \middle| w_1 \cdots w_{j-1}\right)$
- N-gram model

$$P(\boldsymbol{w}) = \prod_{i=1}^{T} P\left(w_j \middle| \boxed{w_{j-N+1}} \cdots w_{j-1}\right)$$

- Log-likelihood:

$$\sum_{i=1,..,N} \log_{\mathrm{b}} P(\boldsymbol{w}^{(i)})$$

- Perplexity (PPL):

$$\mathrm{b}^{\frac{1}{N}\sum_i \log_{\mathrm{b}} P(\boldsymbol{w}^{(i)})}$$

Limitations of N-gram Model

- With increasing order (N) of the N-gram model, the number of possible parameters increases **exponentially**
  - Vocabulary size: V
  - Unigram: $V - 1$ parameters
  - Bigram: $V(V - 1) = V^2 - 1$ parameters
  - N-gram: $V^{N-1} - 1$ parameters
- Require tremendous amount of data to give good estimate on parameters of high-order N-gram models

Neural Network Language Models (NNLM)

- Many word histories are similar (but not exact)
- Project sparse history onto some continuous low-dimensional space
  - i.e., similar histories can be clustered
- Less parameters have to be estimated from the training data

Feedforward NNLM

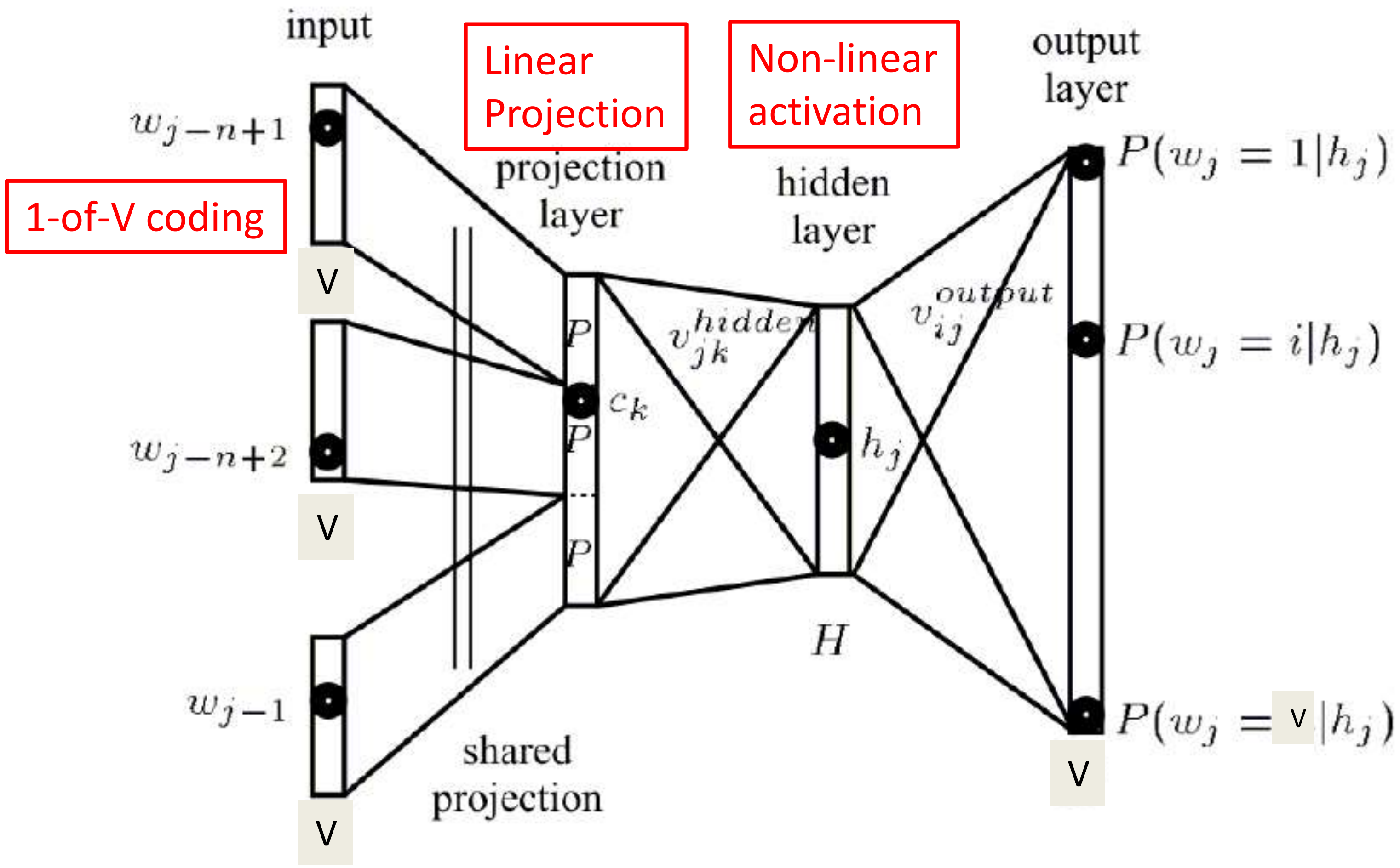

Fig. from Y. Bengio and H. Schwenk.

More effective representation of history?

Recurrent NNLM (RNNLM)

Non-linear activation

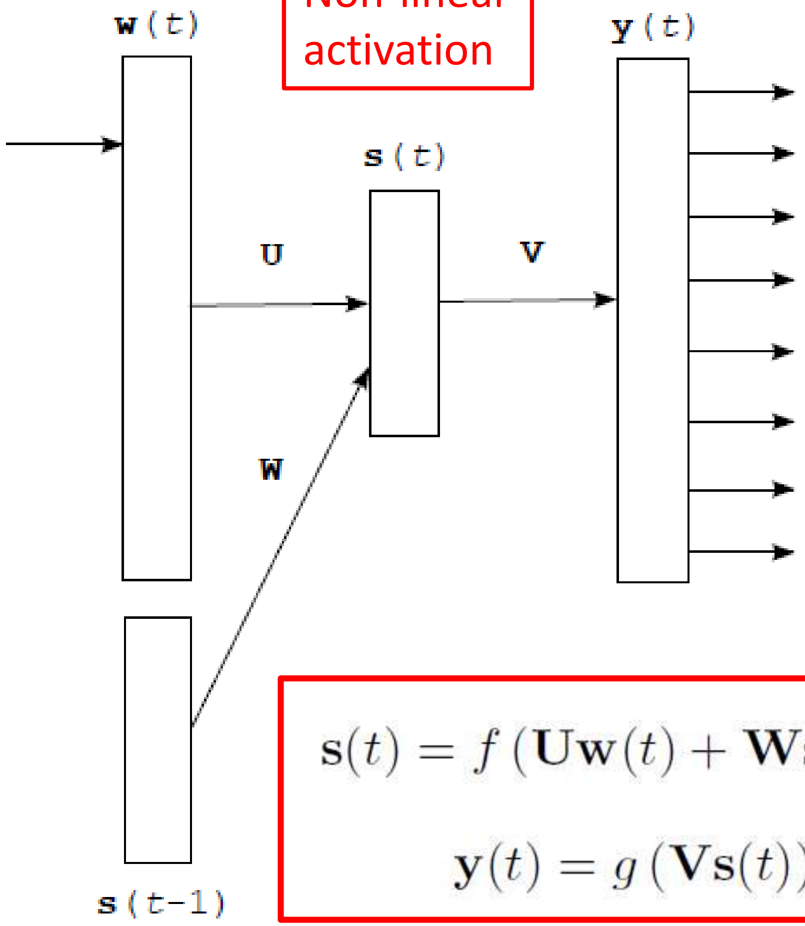

Fig. from T. Mikolov.

- Input layer and output layer have the same dimensionality as the vocabulary
- Hidden layer is smaller (50 – 1000 neurons)
- U, W are the matrices of weights between input and hidden layer
  - U: words
  - W: history states
- V is the matrix of weights between hidden and output layer

$$\boxed{\begin{aligned} \mathbf{s}(t) &= f\left(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1)\right) \\ \mathbf{y}(t) &= g\left(\mathbf{V}\mathbf{s}(t)\right), \end{aligned}}$$

$$\boxed{\begin{aligned} f(z) &= \frac{1}{1+e^{-z}}, \\ g(z_m) &= \frac{e^{z_m}}{\sum_k e^{z_k}} \end{aligned}}$$

Backpropagation Through Time (BPTT)

- The recurrent weights W are updated by unfolding them in time and training the network as a deep feedforward NN.
- The process of propagating errors back through the recurrent weights is called BPTT.

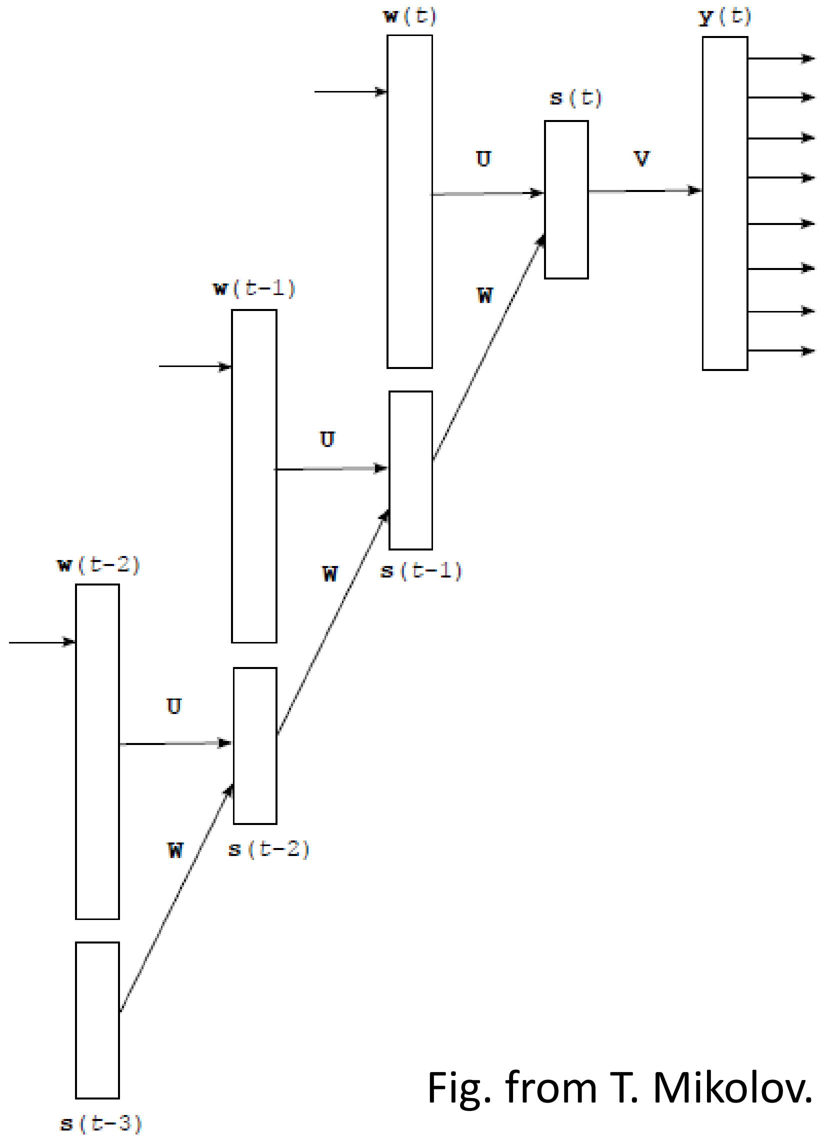

Fig. from T. Mikolov.