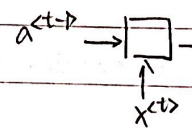


RNN

one-hot encoding

basic RNN

$$\hat{y}^{(t)} = g(W_y a^{(t)} + b_y) \quad \text{softmax}$$

$$a^{(t)} = g(W_{ax} x^{(t)} + W_{aa} a^{(t-1)} + b_a) \quad \text{tanh/relu}$$


vanishing gradient

Gated Recurrent Unit (GRU)

e.g. The cat, which ... , is full ...

\uparrow
c \longrightarrow ?

* simplified version *

$$\tilde{c}^{(t)} = \tanh(W_c [c^{(t-1)}, x^{(t)}] + b_c) \quad \text{what to update given } x^{(t)}, c^{(t-1)}$$

$$I_u = \sigma(W_u [c^{(t-1)}, x^{(t)}] + b_u) \quad \text{whether to update } \tilde{c}^{(t)} \text{ as for } c^{(t)} \text{ (how much)}$$

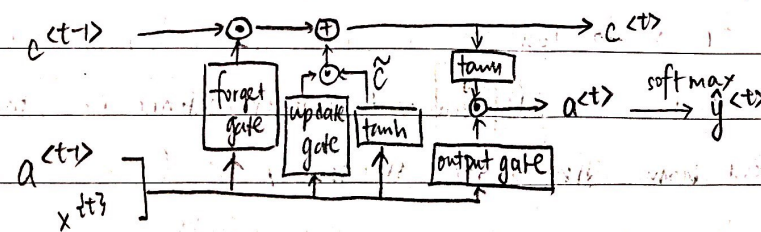
$$c^{(t)} = I_u * \tilde{c}^{(t)} + (1 - I_u) * c^{(t-1)} \quad \hookrightarrow I_f$$

* standard version *

$$\tilde{c}^{(t)} = \tanh(W_c [I_r * c^{(t-1)}, x^{(t)}] + b_c) \quad \text{how relevant is } c^{(t-1)} \text{ to } c^{(t)} \text{ based on } c^{(t-1)} \text{ and } x^{(t)}$$

$$I_r = \sigma(W_r [c^{(t-1)}, x^{(t)}] + b_r)$$

Long Short Term Memory (LSTM)



* variations: e.g. peephole connection, where $c^{(t-1)}$ affect I_o, I_f, I_u

Word Embeddings

analogies, cosine similarity

Skip-gram vs CBOW (continuous bag-of-words): word2vec
target → context context → target

softmax's computational speed problem → hierarchical softmax

negative sampling (adding negative pairs to a matching)

Glove (global vectors for word representation)

* LSAC (latent semantic analysis)

- captures word-word co-occurrence

- minimizes $\sum_i \sum_j f(x_{ij}) (\theta_i^T e_j + b_i + b_j - \log x_{ij})^2$

weighting term symmetric! ↑ ↑ ↓
in case $x_{ij}=0$ ↓ weight weight (e.g. stopwords) co-occurrence probability
this also = 0 $e_w^{(final)} = \frac{e_w + e_w}{2}$

debiasing

Beam search

- in generating sequence, rather than greedy search, consider top N words
and calculate ^{for} potential high-probability phrases

⊕ length normalization

① $P(\dots) \rightarrow \log P(\dots)$ better retain tiny numbers

② $\frac{1}{L^\alpha}$ add penalty for shorter sentence (α is parameter)

- larger beam: width: better result but slower (no optima guaranteed)

BLEU (bilingual evaluation understudy) score

① $\frac{\text{各 } n\text{-gram 在 human translation 中出现的次数}}{\dots \text{ machine } \dots} = P_n$ ② combined: $Bp \cdot e^{\frac{1}{K} \sum P_i}$
brevity penalty