Recurrent Neural Networks (RNNs)

Dr. Benjamin Roth, Nina Poerner

CIS LMU München

Heute

- 9:15 - 10:00: RNN Basics
- 10:15 - 11:45: Übungen: PyTorch, Word2Vec
- Statt Übungsblatt bis nächste Woche durcharbeiten:
  - http://www.deeplearningbook.org/contents/rnn.html (Abschnitte 10.0 - 10.2.1 (inclusive), 10.7, 10.10)
  - LSTM: http://colah.github.io/posts/2015-08-Understanding-LSTMs/
  - GRU: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be/
- Nächste Woche:
  - 9:15 - 10:15: "Journal Club" zu LSTM und GRU
  - 10:30 - 11:45: Intro Keras

Recurrent Neural Networks (RNNs)

- Family of neural networks for processing sequential data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$.
- Sequences of words, characters, video frames, audio frames, ...

Recurrent Neural Networks (RNNs)

- Family of neural networks for processing sequential data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$.
- Sequences of words, characters, video frames, audio frames, ...
- Basic idea: For time step t, compute representation $\mathbf{h}^{(t)}$ from current input $\mathbf{x}^{(t)}$ and previous representation $\mathbf{h}^{(t-1)}$.

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

Recurrent Neural Networks (RNNs)

- Family of neural networks for processing sequential data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$.
- Sequences of words, characters, video frames, audio frames, ...
- Basic idea: For time step t, compute representation $\mathbf{h}^{(t)}$ from current input $\mathbf{x}^{(t)}$ and previous representation $\mathbf{h}^{(t-1)}$.

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

- $\mathbf{x}^{(t)}$ can be embeddings, one-hot, output of some previous layer ...
- **Question:** By recursion, what does $\mathbf{h}^{(t)}$ depend on?

Recurrent Neural Networks (RNNs)

- Family of neural networks for processing sequential data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$.
- Sequences of words, characters, video frames, audio frames, ...
- Basic idea: For time step t, compute representation $\mathbf{h}^{(t)}$ from current input $\mathbf{x}^{(t)}$ and previous representation $\mathbf{h}^{(t-1)}$.

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

- $\mathbf{x}^{(t)}$ can be embeddings, one-hot, output of some previous layer ...
- **Question:** By recursion, what does $\mathbf{h}^{(t)}$ depend on?
  - all previous inputs $\mathbf{x}^{(1)} \dots \mathbf{x}^{(t)}$
  - the initial state $\mathbf{h}^{(0)}$ (typically all-zero, but not necessarily, c.f. encoder-decoder)
  - the parameters θ
- **Question:** How does this compare to an n-gram based model?

Recurrent Neural Networks (RNNs)

- Family of neural networks for processing sequential data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$.
- Sequences of words, characters, video frames, audio frames, ...
- Basic idea: For time step t, compute representation $\mathbf{h}^{(t)}$ from current input $\mathbf{x}^{(t)}$ and previous representation $\mathbf{h}^{(t-1)}$.

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

- $\mathbf{x}^{(t)}$ can be embeddings, one-hot, output of some previous layer ...
- **Question:** By recursion, what does $\mathbf{h}^{(t)}$ depend on?
  - all previous inputs $\mathbf{x}^{(1)} \dots \mathbf{x}^{(t)}$
  - the initial state $\mathbf{h}^{(0)}$ (typically all-zero, but not necessarily, c.f. encoder-decoder)
  - the parameters θ
- **Question:** How does this compare to an n-gram based model?
  - N-gram based models have limited memory, the RNN has theoretically (!) unlimited memory

Parameter Sharing

- Going from a time step $t-1$ to t is parameterized by the same parameters θ for all t!

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

- **Question:** Why is parameter sharing a good idea?

Parameter Sharing

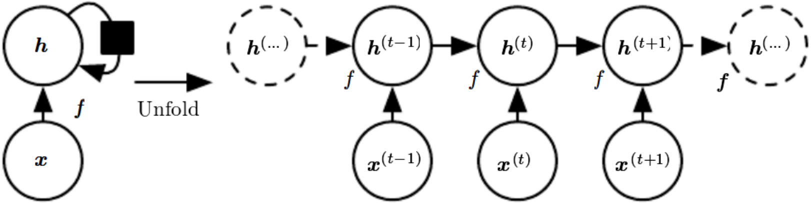
- Going from a time step $t-1$ to t is parameterized by the same parameters θ for all t!

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta)$$

- **Question:** Why is parameter sharing a good idea?
  - Fewer parameters
  - Can learn to detect features regardless of their position
    - "i went to nepal in 2009" vs. "in 2009 i went to nepal"
  - Can generalize to longer sequences than were seen in training

Graphical Notation: Unrolling

- Compact notation (left):
  - All time steps conflated.
  - ■ indicates *"delay"* of 1 time unit.
- Unrolled notation (right):
  - Like a very deep feed-forward NN with parameter sharing across layers



Source: Goodfellow et al.: Deep Learning.

Any questions so far?

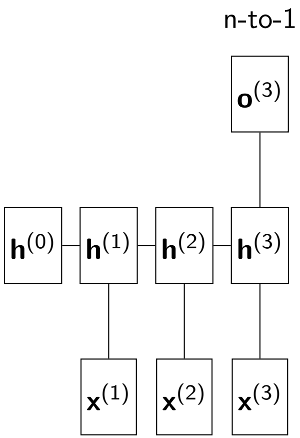
RNN: Output

- The output at time t is typically computed from the hidden representation at time t:

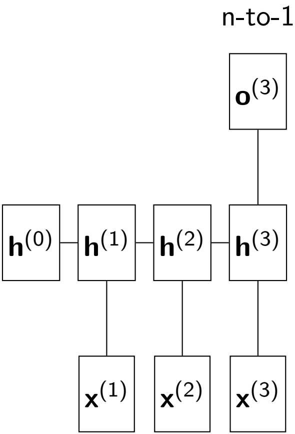
$$\mathbf{o}^{(t)} = f(\mathbf{h}^{(t)}; \theta_o)$$

- Typically a linear transformation: $\mathbf{o}^{(t)} = \theta_o^T \mathbf{h}^{(t)}$
- Some RNNs compute $\mathbf{o}^{(t)}$ at every time step, others only at the last time step $\mathbf{o}^{(T)}$

RNN: Output

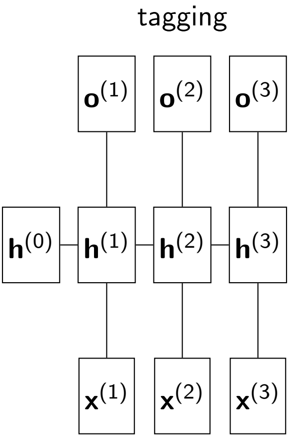

n-to-1
$\mathbf{o}^{(3)}$
$\mathbf{h}^{(0)}$
$\mathbf{h}^{(1)}$
$\mathbf{h}^{(2)}$
$\mathbf{h}^{(3)}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(3)}$

RNN: Output



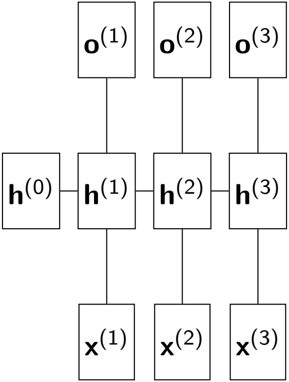
Sentiment polarity, topic classification, grammaticality ...

RNN: Output


tagging
$\mathbf{o}^{(1)}$
$\mathbf{o}^{(2)}$
$\mathbf{o}^{(3)}$
$\mathbf{h}^{(0)}$
$\mathbf{h}^{(1)}$
$\mathbf{h}^{(2)}$
$\mathbf{h}^{(3)}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(3)}$

RNN: Output

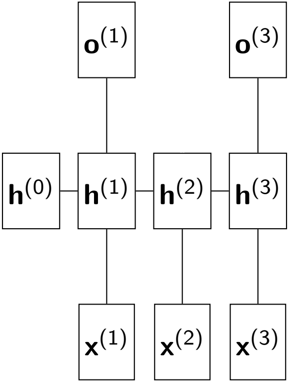
tagging



POS tagging, NER tagging, Language Model ...

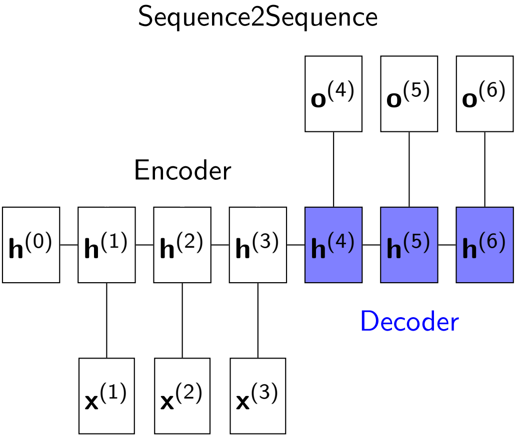
RNN: Output

tagging (selective)

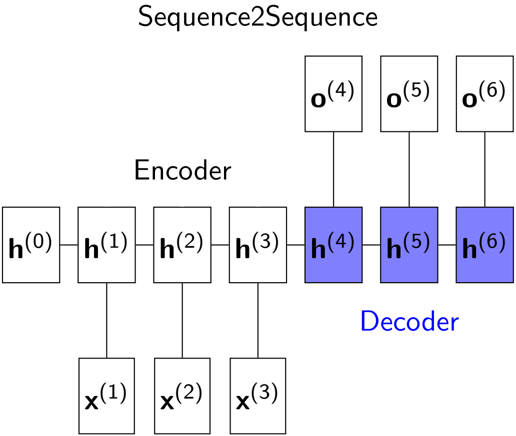


POS tagging, NER tagging, Language Model ...

RNN: Output


Sequence2Sequence
o(4)
o(5)
o(6)
Encoder
h(0)
h(1)
h(2)
h(3)
h(4)
h(5)
h(6)
Decoder
x(1)
x(2)
x(3)

RNN: Output



Machine Translation, Summarization, Image captioning (encoder CNN) ...

Any questions so far?

RNN: Loss Function

- Loss function:
  - Several time steps: $\mathcal{L}(y^{(1)}, \ldots y^{(T)}; \mathbf{o}^{(1)} \ldots \mathbf{o}^{(T)})$
  - Last time step: $\mathcal{L}(y; \mathbf{o}^{(T)})$

RNN: Loss Function

- Loss function:
  - Several time steps: $\mathcal{L}(y^{(1)}, \ldots y^{(T)}; \mathbf{o}^{(1)} \ldots \mathbf{o}^{(T)})$
  - Last time step: $\mathcal{L}(y; \mathbf{o}^{(T)})$
- Example: POS Tagging
  - Output $\mathbf{o}^{(t)}$ is predicted distribution over POS tags
    - ★ $\mathbf{o}^{(t)} = P(\text{tag} = ?|\mathbf{h}^{(t)})$
    - ★ Typically: $\mathbf{o}^{(t)} = \mathrm{softmax}(\theta_o^T \mathbf{h}^{(t)})$

RNN: Loss Function

- Loss function:
  - Several time steps: $\mathcal{L}(y^{(1)}, \dots y^{(T)}; \mathbf{o}^{(1)} \dots \mathbf{o}^{(T)})$
  - Last time step: $\mathcal{L}(y; \mathbf{o}^{(T)})$
- Example: POS Tagging
  - Output $\mathbf{o}^{(t)}$ is predicted distribution over POS tags
    - ★ $\mathbf{o}^{(t)} = P(\text{tag} = ?|\mathbf{h}^{(t)})$
    - ★ Typically: $\mathbf{o}^{(t)} = \mathrm{softmax}(\theta_o^T \mathbf{h}^{(t)})$
  - Loss at time t: negative log-likelihood (NLL) of true label $y^{(t)}$

$$\mathcal{L}^{(t)} = -\log P(\text{tag} = y^{(t)}|\mathbf{h}^{(t)}; \theta_o)$$

RNN: Loss Function

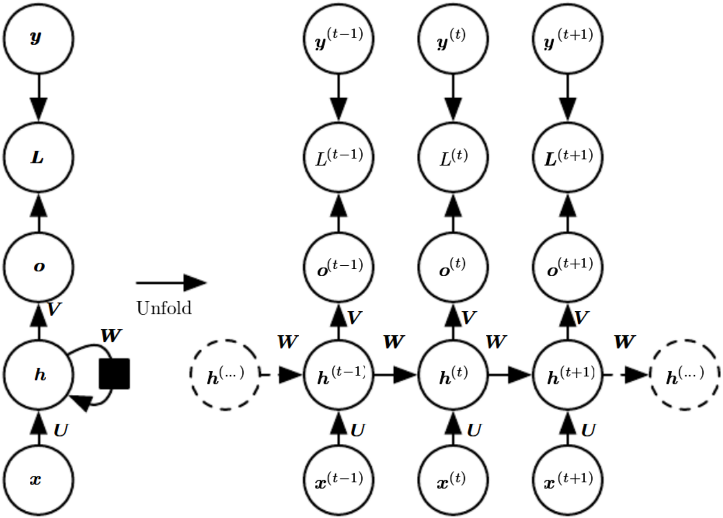
- Loss function:
  - Several time steps: $\mathcal{L}(y^{(1)}, \ldots y^{(T)}; \mathbf{o}^{(1)} \ldots \mathbf{o}^{(T)})$
  - Last time step: $\mathcal{L}(y; \mathbf{o}^{(T)})$
- Example: POS Tagging
  - Output $\mathbf{o}^{(t)}$ is predicted distribution over POS tags
    - ★ $\mathbf{o}^{(t)} = P(\text{tag} = ?|\mathbf{h}^{(t)})$
    - ★ Typically: $\mathbf{o}^{(t)} = \mathrm{softmax}(\theta_o^T \mathbf{h}^{(t)})$
  - Loss at time t: negative log-likelihood (NLL) of true label $y^{(t)}$

$$\mathcal{L}^{(t)} = -\log P(\text{tag} = y^{(t)}|\mathbf{h}^{(t)}; \theta_o)$$

  - Overall Loss for all time steps:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}^{(t)}$$

Graphical Notation: Including Output and Loss Function



Source: Goodfellow et al.: Deep Learning.

Any questions so far?

Backpropagation through time

- We have calculated loss $\mathcal{L}^{(T)}$ at time step T and want to update parameters θ_i with gradient descent.

Backpropagation through time

- We have calculated loss $\mathcal{L}^{(T)}$ at time step T and want to update parameters θ_i with gradient descent.
- For now, imagine that we have time step-specific "dummy"-parameters $\theta_i^{(t)}$, which are identical copies of θ_i
- $\rightarrow$ the unrolled RNN looks like a feed-forward-neural-network!
- $\rightarrow$ we can calculate $\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i^{(t)}}$ using standard backpropagation

Backpropagation through time

- We have calculated loss $\mathcal{L}^{(T)}$ at time step T and want to update parameters θ_i with gradient descent.
- For now, imagine that we have time step-specific "dummy"-parameters $\theta_i^{(t)}$, which are identical copies of θ_i
- $\rightarrow$ the unrolled RNN looks like a feed-forward-neural-network!
- $\rightarrow$ we can calculate $\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i^{(t)}}$ using standard backpropagation
- To calculate $\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i}$, add up the "dummy" gradients:

$$\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i^{(t)}}$$

Truncated backpropagation through time

- Simple idea: Stop backpropagation through time after k time steps

$$\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i} = \sum_{t=T-k}^{T} \frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i^{(t)}}$$

- **Question:** What are advantages and disadvantages?

Truncated backpropagation through time

- Simple idea: Stop backpropagation through time after k time steps

$$\frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i} = \sum_{t=T-k}^{T} \frac{\partial \mathcal{L}^{(T)}}{\partial \theta_i^{(t)}}$$

- **Question:** What are advantages and disadvantages?
  - Advantage: Faster and parallelizable
  - Disadvantage: If k is too small, long-range dependencies are hard to learn

Any questions so far?

Vanilla RNN

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta) = \tanh(\mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(h-1)} + \mathbf{b})$$

$$\theta = \{\mathbf{W}, \mathbf{U}, \mathbf{b}\}$$

- **W**: Hidden-to-hidden
- **U**: Input-to-hidden
- **b**: Bias term

Vanilla RNN

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta) = \tanh(\mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(h-1)} + \mathbf{b})$$

$$\theta = \{\mathbf{W}, \mathbf{U}, \mathbf{b}\}$$

- $\mathbf{W}$: Hidden-to-hidden
- $\mathbf{U}$: Input-to-hidden
- $\mathbf{b}$: Bias term
- Vanilla RNN in keras:

```
vanilla = SimpleRNN(units=10, use_bias = True)
vanilla.build(input_shape = (None, None, 30))
print([weight.shape for weight in vanilla.get_weights()])
[(30, 10), (10, 10), (10,)]
```

- **Question:** Which shape belongs to which weight?

Bidirectional RNNs

- Conceptually: Two RNNs that run in opposite directions over the same input
- Typically, each RNN has its own set of parameters
- Results in two sequences of hidden vectors: $\overrightarrow{\mathbf{h}}^{(1)} \dots \overrightarrow{\mathbf{h}}^{(T)}$, $\overleftarrow{\mathbf{h}}^{(1)} \dots \overleftarrow{\mathbf{h}}^{(T)}$

- Before being passed to downstream layers, $\overset{\rightarrow}{\mathbf{h}}$ and $\overset{\leftarrow}{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overset{\rightarrow}{\mathbf{h}}) + \dim(\overset{\leftarrow}{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?

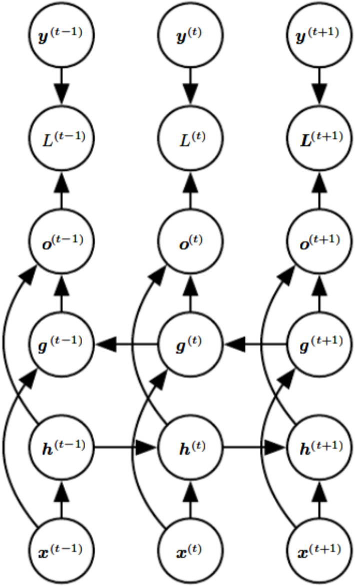
- Before being passed to downstream layers, $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overrightarrow{\mathbf{h}}) + \dim(\overleftarrow{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?
  - $\mathbf{h} = \overrightarrow{\mathbf{h}}^{(T)} || \overleftarrow{\mathbf{h}}^{(1)}$
  - Because these are the vectors that have "read" the entire sequence

- Before being passed to downstream layers, $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overrightarrow{\mathbf{h}}) + \dim(\overleftarrow{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?
  - $\mathbf{h} = \overrightarrow{\mathbf{h}}^{(T)} || \overleftarrow{\mathbf{h}}^{(1)}$
  - Because these are the vectors that have "read" the entire sequence
- **Question:** Which hidden vectors should we concatenate in sequence tagging (e.g., to predict POS for word w_t)?

- Before being passed to downstream layers, $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overrightarrow{\mathbf{h}}) + \dim(\overleftarrow{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?
  - $\mathbf{h} = \overrightarrow{\mathbf{h}}^{(T)} || \overleftarrow{\mathbf{h}}^{(1)}$
  - Because these are the vectors that have "read" the entire sequence
- **Question:** Which hidden vectors should we concatenate in sequence tagging (e.g., to predict POS for word w_t)?
  - $\mathbf{h}^{(t)} = \overrightarrow{\mathbf{h}}^{(t)} || \overleftarrow{\mathbf{h}}^{(t)}$
  - Left context, right context (including t)

- Before being passed to downstream layers, $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overrightarrow{\mathbf{h}}) + \dim(\overleftarrow{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?
  - $\mathbf{h} = \overrightarrow{\mathbf{h}}^{(T)} || \overleftarrow{\mathbf{h}}^{(1)}$
  - Because these are the vectors that have "read" the entire sequence
- **Question:** Which hidden vectors should we concatenate in sequence tagging (e.g., to predict POS for word w_t)?
  - $\mathbf{h}^{(t)} = \overrightarrow{\mathbf{h}}^{(t)} || \overleftarrow{\mathbf{h}}^{(t)}$
  - Left context, right context (including t)
- **Question:** Which hidden vectors should we concatenate to predict symbol w_t (i.e., in a bidirectional language model)?

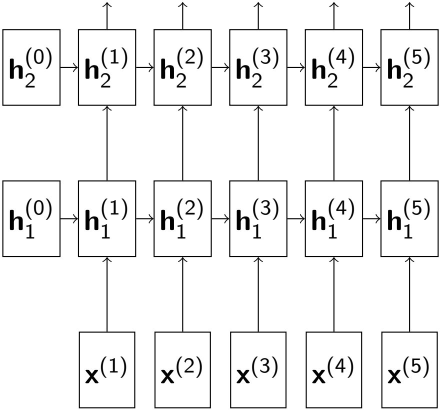
- Before being passed to downstream layers, $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are typically concatenated into one representation $\mathbf{h}$, s.t. $\dim(\mathbf{h}) = \dim(\overrightarrow{\mathbf{h}}) + \dim(\overleftarrow{\mathbf{h}})$.
- **Question:** Which hidden vectors should we concatenate if we want to compute a single output (e.g., predict sentiment of sentence)?
  - $\mathbf{h} = \overrightarrow{\mathbf{h}}^{(T)} || \overleftarrow{\mathbf{h}}^{(1)}$
  - Because these are the vectors that have "read" the entire sequence
- **Question:** Which hidden vectors should we concatenate in sequence tagging (e.g., to predict POS for word w_t)?
  - $\mathbf{h}^{(t)} = \overrightarrow{\mathbf{h}}^{(t)} || \overleftarrow{\mathbf{h}}^{(t)}$
  - Left context, right context (including t)
- **Question:** Which hidden vectors should we concatenate to predict symbol w_t (i.e., in a bidirectional language model)?
  - $\overrightarrow{\mathbf{h}}^{(t-1)} || \overleftarrow{\mathbf{h}}^{(t+1)}$
  - Left context, right context (excluding t)



Bidirectional RNN for sequence tagging. Source: Goodfellow et al.: Deep Learning.

Multi-Layer RNNs

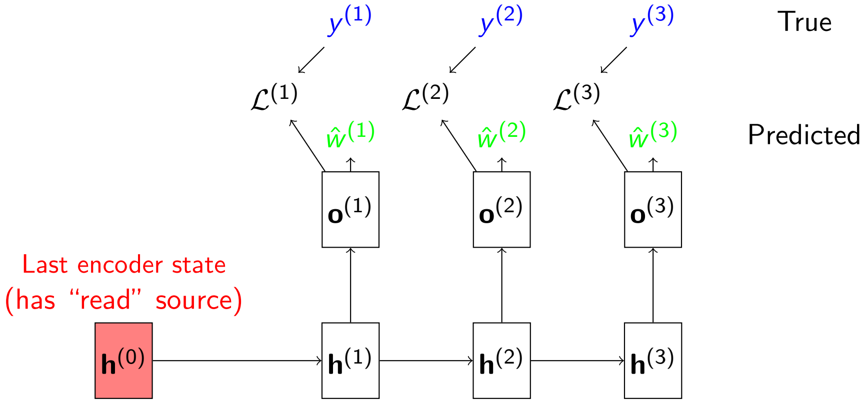
- Conceptually: A stack of L RNNs, such that $\mathbf{x}_l^{(t)} = \mathbf{h}_{l-1}^{(t)}$.



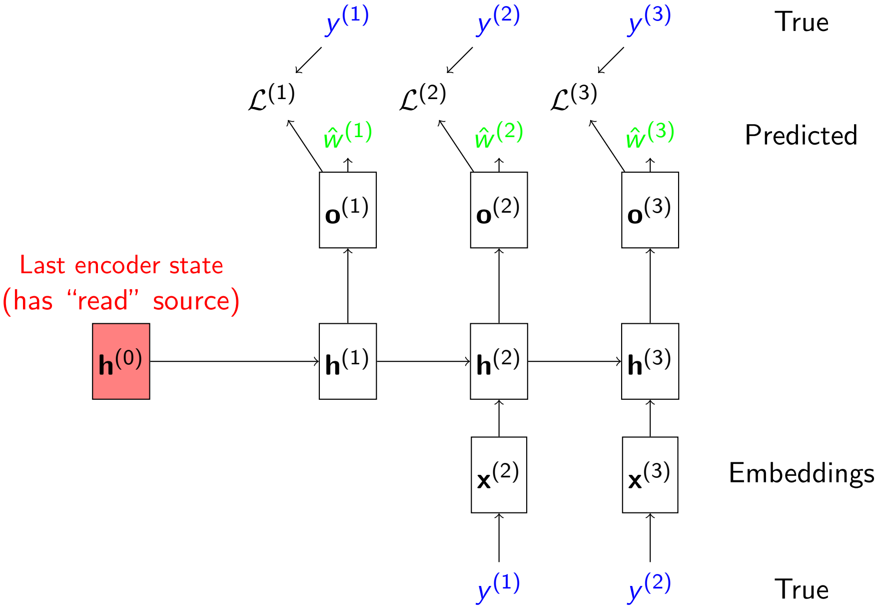
Feeding outputs back

- What do we do if the input sequence $\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}$ is only given at training time, but not at test time?
- Examples: Machine Translation decoder, (generative) language model

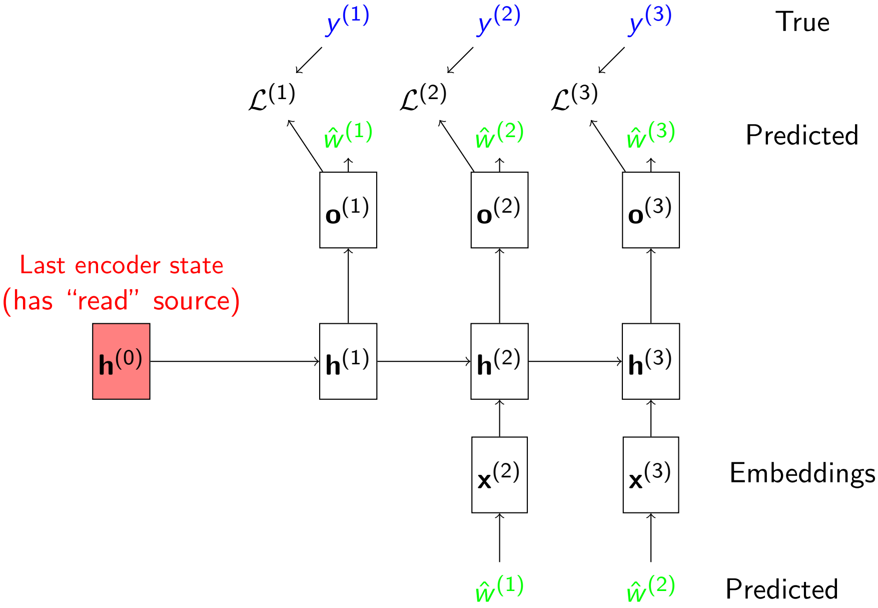
Example: Machine Translation



Example: Machine Translation



Example: Machine Translation



Oracle

- Give Neural Network a signal that it will not have at test time
- Can be useful during training (e.g., mix oracle and predicted signal when training a generative language model)
- Can be used to establish upper bounds of modules
  - Example: How much better do Neural MT systems become when they take the translation of the previous sentence into account?
  - If we don't see improvements, this could be because
    - ★ the previous sentence contains no useful information in general
    - ★ the translation of the previous sentence was not good enough to have a positive effect
  - $\rightarrow$ provide gold translation of previous sentence as oracle to find upper bound

Gated RNNs: Teaser

- Vanilla RNNs are not frequently used, because
  - Vanishing/exploding gradients make them difficult to train
  - They tend to forget past information quickly
- Instead: LSTM, GRU, ... (next week!)