Recurrent Neural Networks, I

LING 575K Deep Learning for NLP

Shane Steinert-Threlkeld

April 19 2021

Today's Plan

- Last time:
  - Deep Averaging Networks for text classification
- Neural Probabilistic Language Model
- Additional Training Notes
  - Regularization
  - Early stopping
  - Hyper-parameter searching
- Intro to *Recurrent* Neural Networks

Announcements

- HW2 reference code (and symlinks from hw3) available now
- HW3 tests: hw3/test_all.py. NB: necessary, but not sufficient, to check correctness of your code. `pytest test_all.py`, from your directory, with environment activated.
- Implementing ops in edugrad:
    - You can use any numpy operations you want; goal it to understand forward/backward API
    - https://github.com/shanest/edugrad
    - Log: base e, don't need to do special handling of bad input arguments (like 0)
- Edugrad is installed in the course conda environment, so be sure to activate it
- $f(x) = x^2 \times 3x$ and static computation graphs

Decorators

- @tensor_op in edugrad code: what is this??
- Example of a decorator
  - Design pattern to extend an object with more functionality
  - Decorators *wrap* their arguments, add features (e.g. registering in a central DB)
- In python, syntactic sugar:
- Canonical examples:
  - @classmethod
  - @staticmethod

```
@my_decorator
def fn(...):


def fn(...):

fn = my_decorator(fn)
```

Decorator Demo

```python
def printer(method, *args):
    def fn(*args):
        output = method(*args)
        print(f"Output: {output}")
    return fn

@printer
def add(a, b):
    return a + b

add(1, 2) # prints "Output: 3"
```

Unordered Models in the Large LM Era

- Last time: “Deep Unordered Composition Rivals Syntactic Methods for Text Classification” —2015
- Brand new paper:

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Montreal Institute of Learning Algorithms

{koustuvs,adinawilliams,dkiela}@fb.com

Abstract

A possible explanation for the impressive performance of masked language model (MLM) pre-training is that such models have learned to represent the syntactic structures prevalent in classical NLP pipelines. In this paper, we propose a different explanation: MLMs succeed on downstream tasks almost entirely due to their ability to model higher-order word co-occurrence statistics. To demonstrate this, we pre-train MLMs on sentences with randomly shuffled word order, and show that

NLP pipeline” (Tenney et al., 2019), suggesting that it has learned “the kind of abstractions that we intuitively believe are important for representing natural language” rather than “simply modeling complex co-occurrence statistics” (ibid., p. 1).

In this work, we try to uncover how much of MLM’s success comes from simple distributional information, as opposed to “the types of syntactic and semantic abstractions traditionally believed necessary for language processing” (Tenney et al., 2019; Manning et al., 2020). We disentangle these

Unordered Models in the Large LM Era

- Last time: “Deep Unordered Composition Rivals Syntactic Methods for Text Classification” —2015

- Brand new paper:

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Montreal Institute of Learning Algorithms

{koustuvs,adinawilliams,dkiela}@fb.com

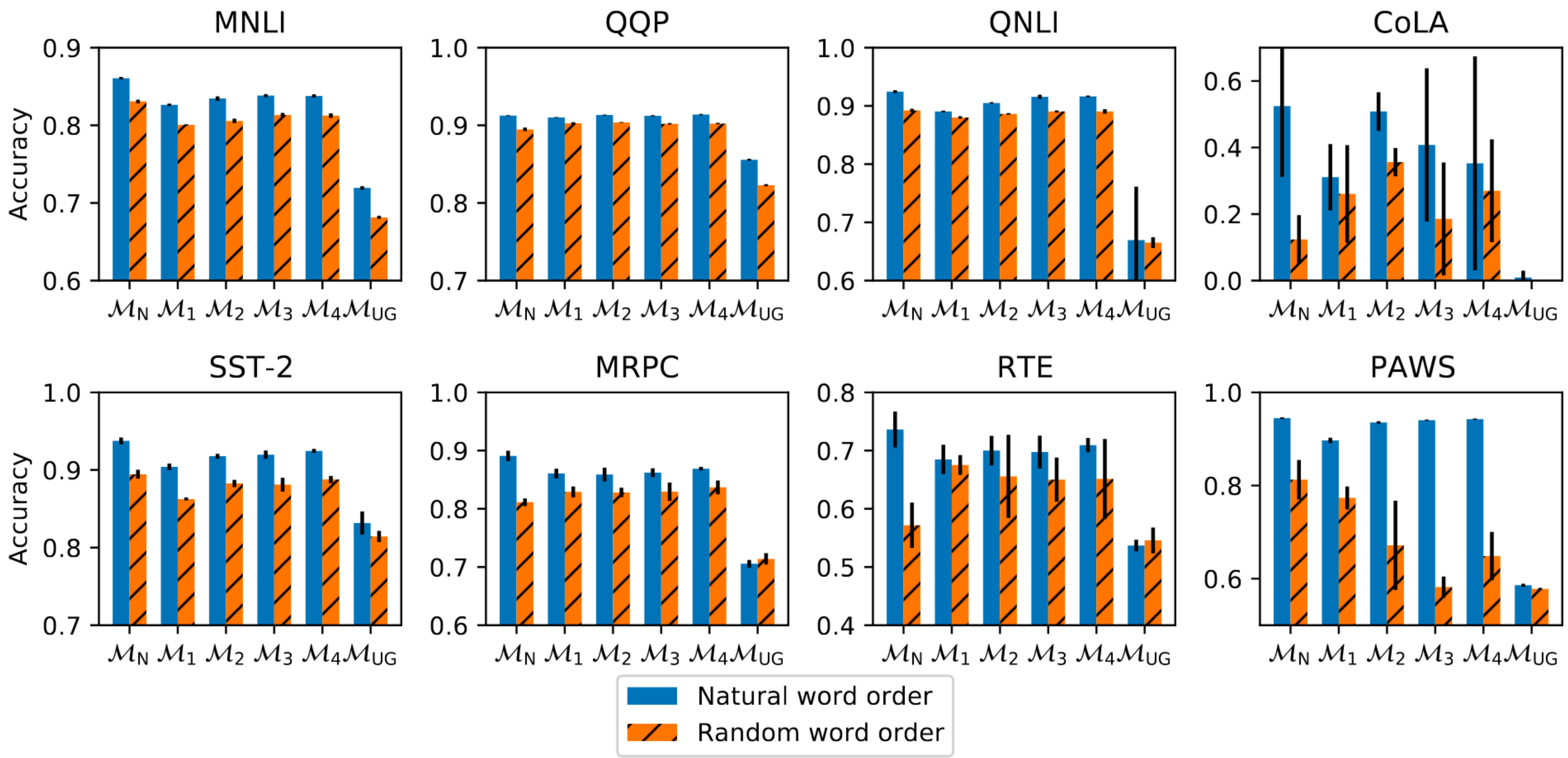
Abstract

A possible explanation for the impressive performance of masked language model (MLM) pre-training is that such models have learned to represent the syntactic structures prevalent in classical NLP pipelines. In this paper, we propose a different explanation: MLMs succeed on downstream tasks almost entirely due to their ability to model higher-order word co-occurrence statistics. To demonstrate this, we pre-train MLMs on sentences with randomly shuffled word order, and show that

NLP pipeline” (Tenney et al., 2019), suggesting that it has learned “the kind of abstractions that we intuitively believe are important for representing natural language” rather than “simply modeling complex co-occurrence statistics” (ibid., p. 1).

In this work, we try to uncover how much of MLM’s success comes from simple distributional information, as opposed to “the types of syntactic and semantic abstractions traditionally believed necessary for language processing” (Tenney et al., 2019; Manning et al., 2020). We disentangle these

Unordered Models in the Large LM Era



Unordered Models in the Large LM Era

- “We observed overwhelmingly that MLM’s success is most likely **not** [emphasis added] due to its ability to discover syntactic and semantic mechanisms necessary for a traditional language processing pipeline. Instead, our experiments suggest that MLM’s success can be mostly explained by it having learned higher-order distributional statistics that make for a useful prior for subsequent fine-tuning.”

Recurrent Neural Networks

RNNs: high-level

RNNs: high-level

- Feed-forward networks: fixed-size input, fixed-size output
  - Previous classifier: average embeddings of words
  - Previous LM: *n*-gram assumption (i.e. fixed-size context of word embeddings)

RNNs: high-level

- Feed-forward networks: fixed-size input, fixed-size output
  - Previous classifier: average embeddings of words
  - Previous LM: *n*-gram assumption (i.e. fixed-size context of word embeddings)
- RNNs process *sequences* of vectors
  - Maintaining “hidden” state
  - Applying the same operation at each step

RNNs: high-level

- Feed-forward networks: fixed-size input, fixed-size output
  - Previous classifier: average embeddings of words
  - Previous LM: *n*-gram assumption (i.e. fixed-size context of word embeddings)
- RNNs process *sequences* of vectors
  - Maintaining "hidden" state
  - Applying the same operation at each step
- Different RNNs:
  - Different operations at each step
  - Operation also called "recurrent cell"
  - Other architectural considerations (e.g. depth; bidirectionally)

Long-distance dependencies, I: number

- Language modeling (fill-in-the-blank)
  - The keys ____
  - The keys on the table ____
  - The keys next to the book on top of the table ____
- To get the number on the verb, need to look at the subject, which can be very far away
  - And number can disagree with linearly-close nouns

Selectional Restrictions

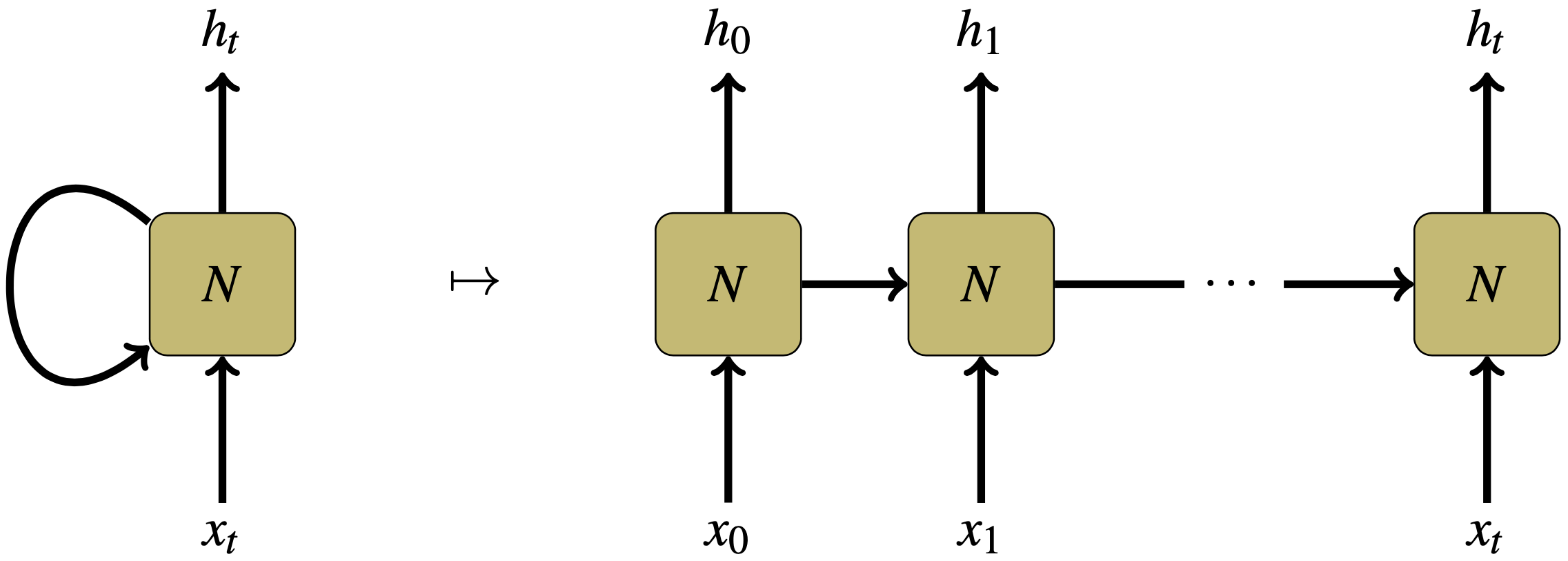
- The **family** moved from the city because they wanted a larger _____.
- The **team** moved from the city because they wanted a larger _____.

Selectional Restrictions

- The **family** moved from the city because they wanted a larger **house**.
- The **team** moved from the city because they wanted a larger **market**.

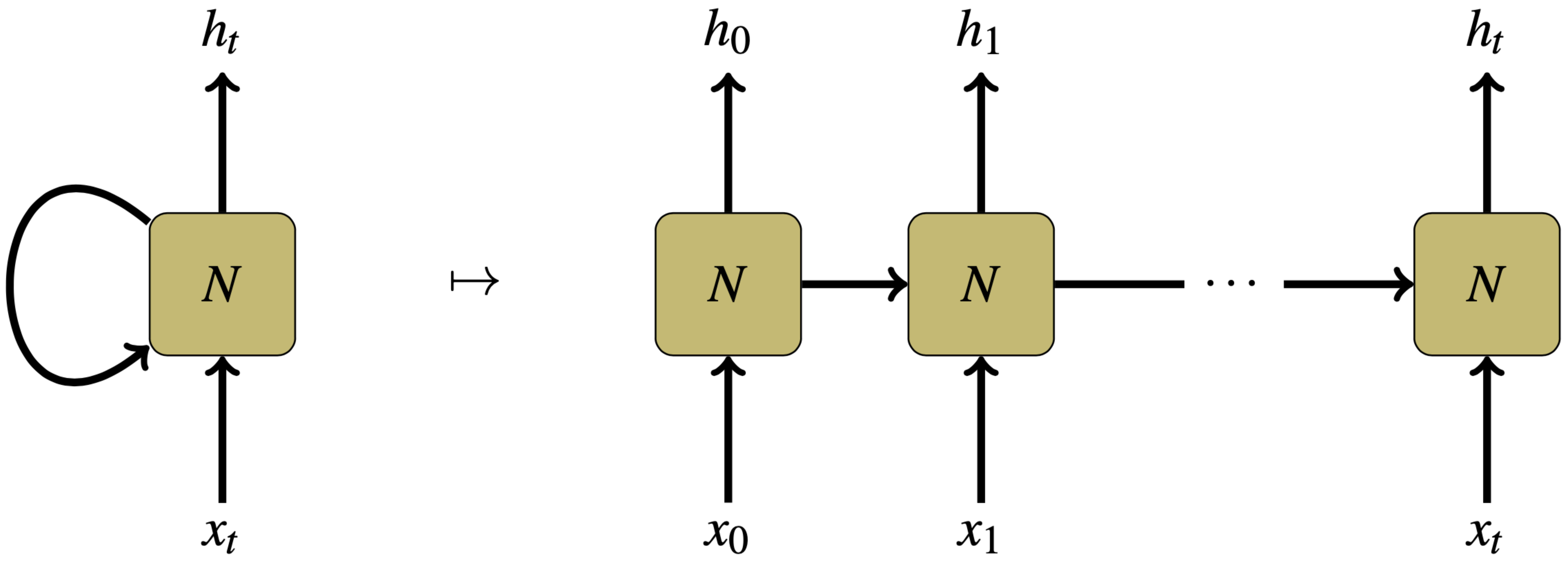
- Need models that can capture long-range dependencies like this.
- N-gram (whether count-based or neural) cannot. E.g., with n=4:
  - P(word | “they wanted a larger”)

RNNs



Steinert-Threlkeld and Szymanik 2019; Olah 2015

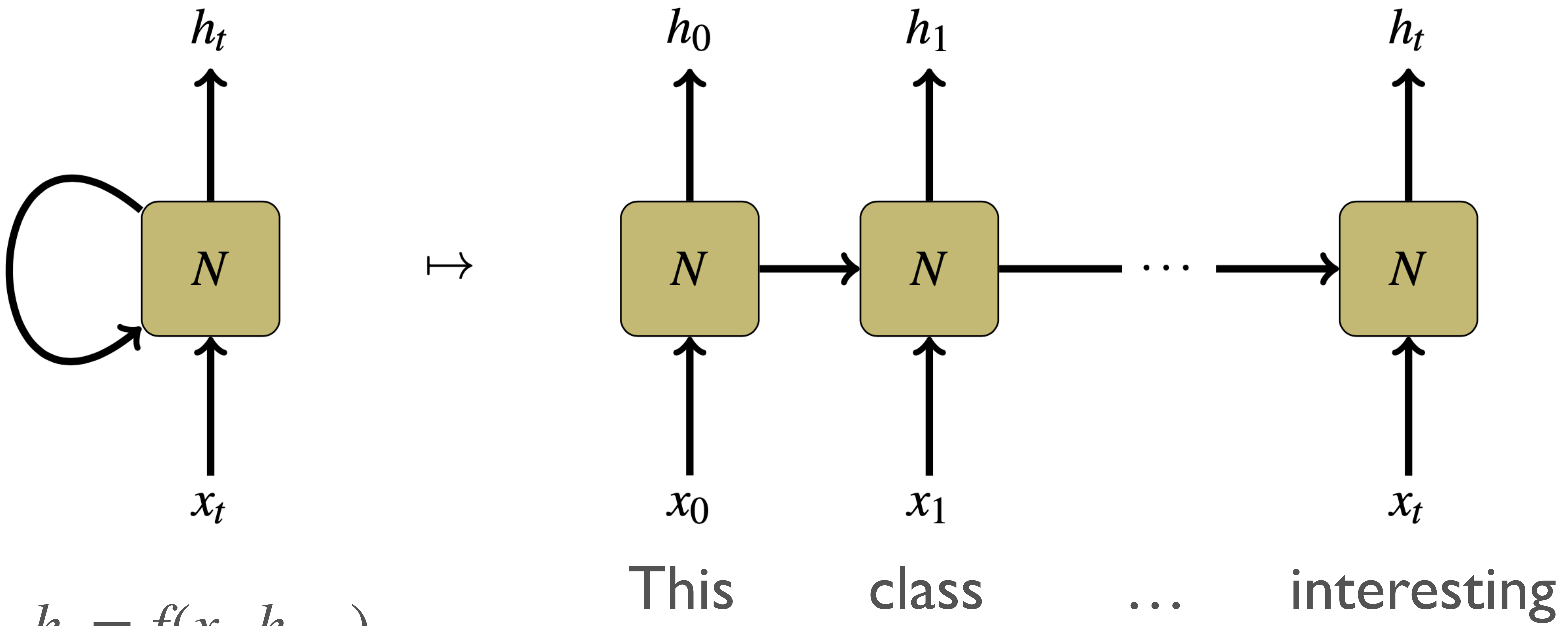
RNNs



$$h_t = f(x_t, h_{t-1})$$

Steinert-Threlkeld and Szymanik 2019; Olah 2015

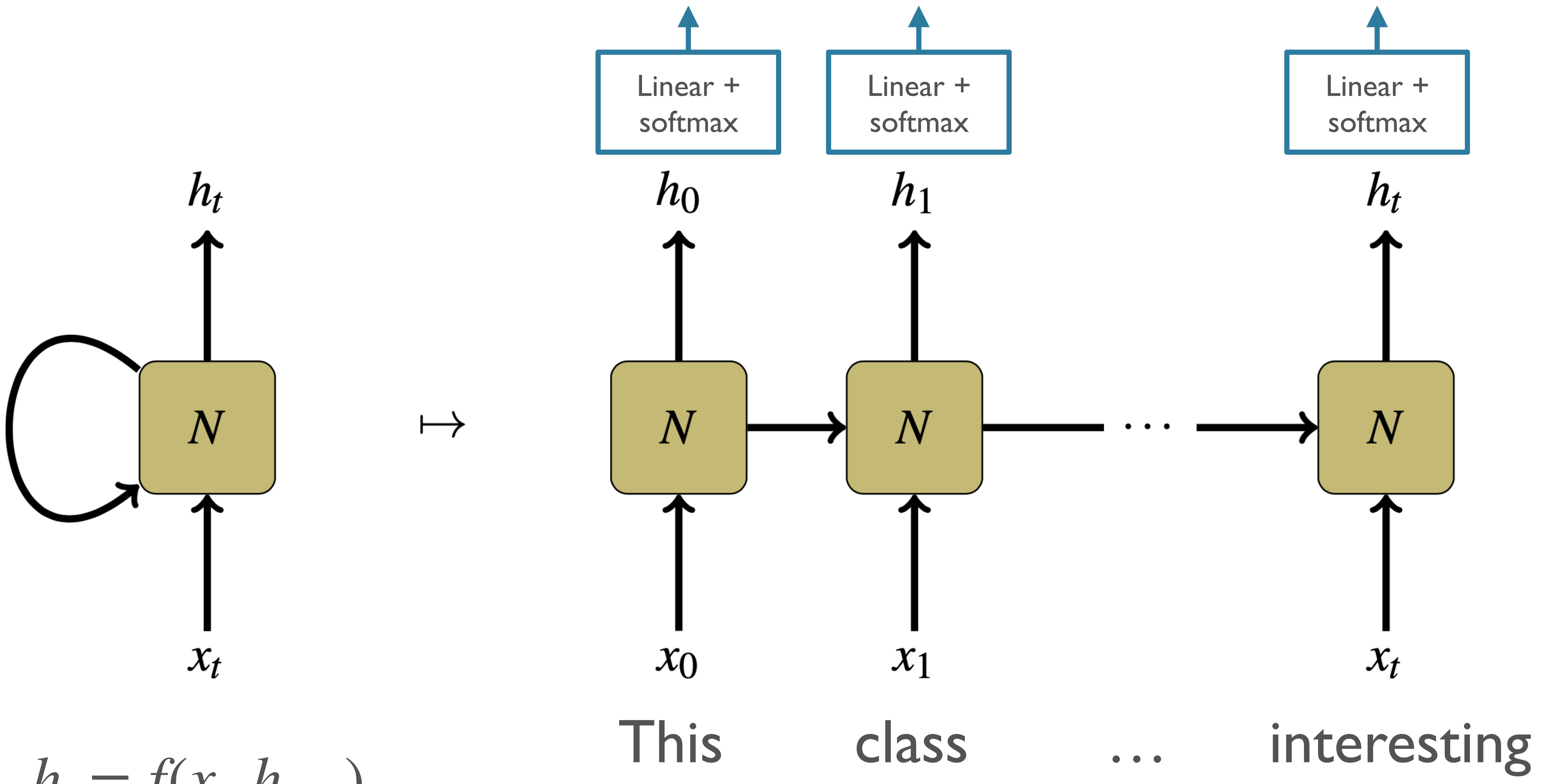
RNNs



$$h_t = f(x_t, h_{t-1})$$

Steinert-Threlkeld and Szymanik 2019; Olah 2015

RNNs



$$h_t = f(x_t, h_{t-1})$$

Steinert-Threlkeld and Szymanik 2019; Olah 2015

Simple / Vanilla / Elman RNNs

- Same kind of feed-forward computation we've been studying, but:
  - x_t: sequence element at time *t*
  - h_{t-1}: hidden state *of the model* at previous time *t-1*

Simple / Vanilla / Elman RNNs

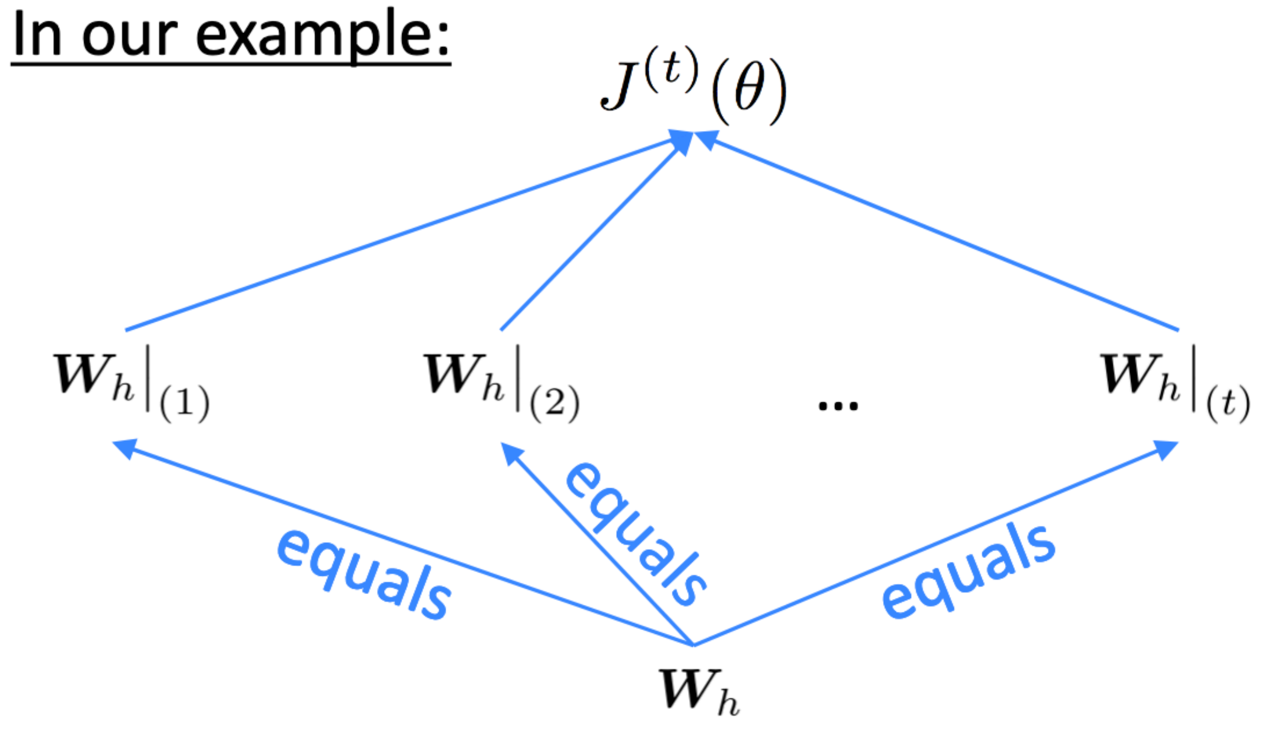
- Same kind of feed-forward computation we've been studying, but:
  - x_t: sequence element at time *t*
  - h_{t-1}: hidden state *of the model* at previous time *t-1*

Simple/"Vanilla" RNN: $$h_t = \tanh(x_t W_x + h_{t-1} W_h + b)$$

Training: BPTT

- Backpropagation Through Time
- "Unroll" the network across time-steps
- Apply backprop to the "wide" network
  - Each cell has the *same* parameters
  - Gradients sum across time-steps
    - Multi-variable chain rule

<u>In our example:</u>



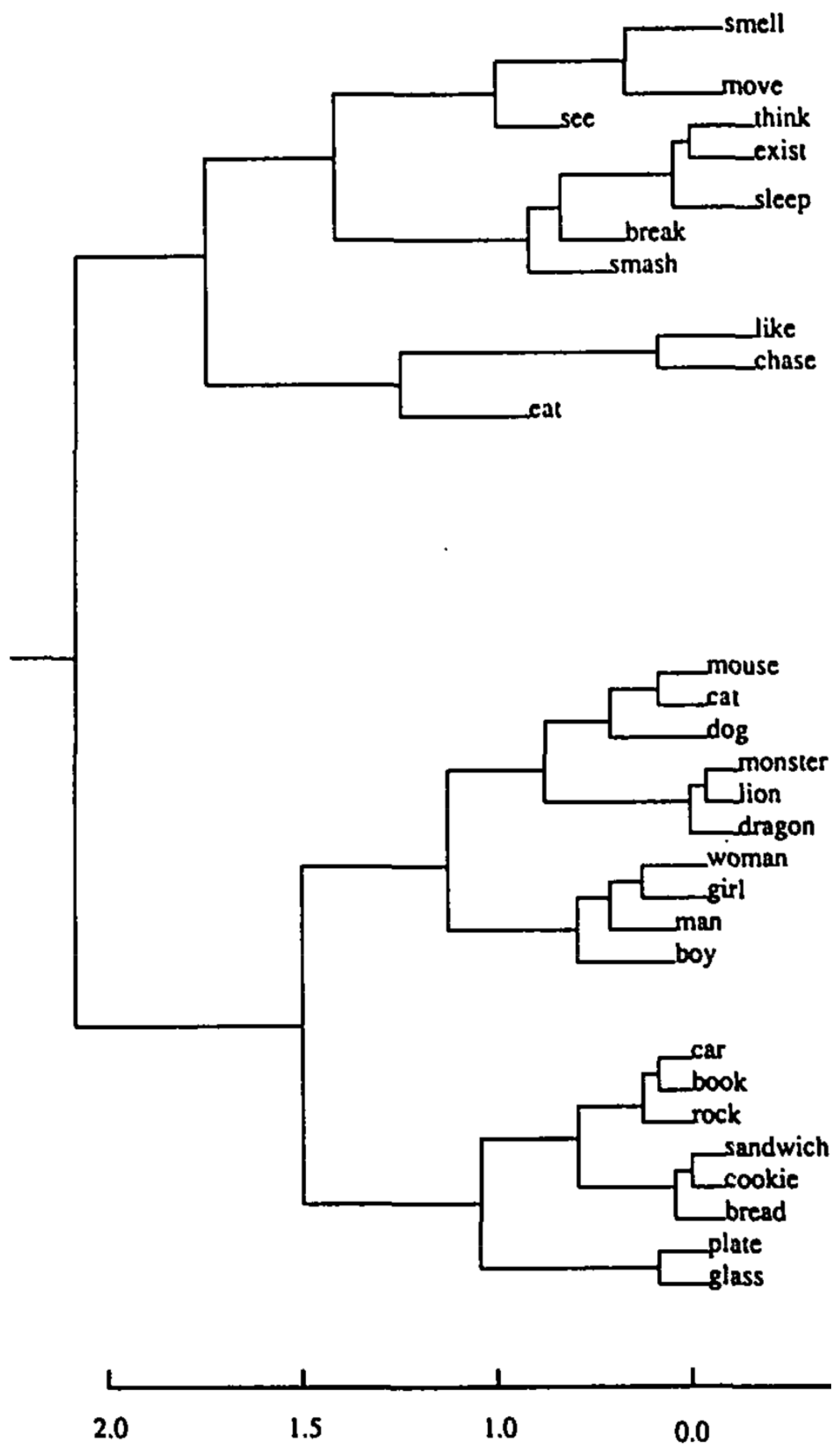
<u>Apply the multivariable chain rule:</u>

$$\frac{\partial J^{(t)}}{\partial \boldsymbol{W}_h} = \sum_{i=1}^{t} \frac{\partial J^{(t)}}{\partial \boldsymbol{W}_h}\bigg|_{(i)} \overbrace{\boxed{\frac{\partial \boldsymbol{W}_h\big|_{(i)}}{\partial \boldsymbol{W}_h}}}^{=1}$$

$$= \sum_{i=1}^{t} \frac{\partial J^{(t)}}{\partial \boldsymbol{W}_h}\bigg|_{(i)}$$

Power of RNNs

Hierarchical clustering of Vanilla RNN hidden states trained as LM on synthetic data:

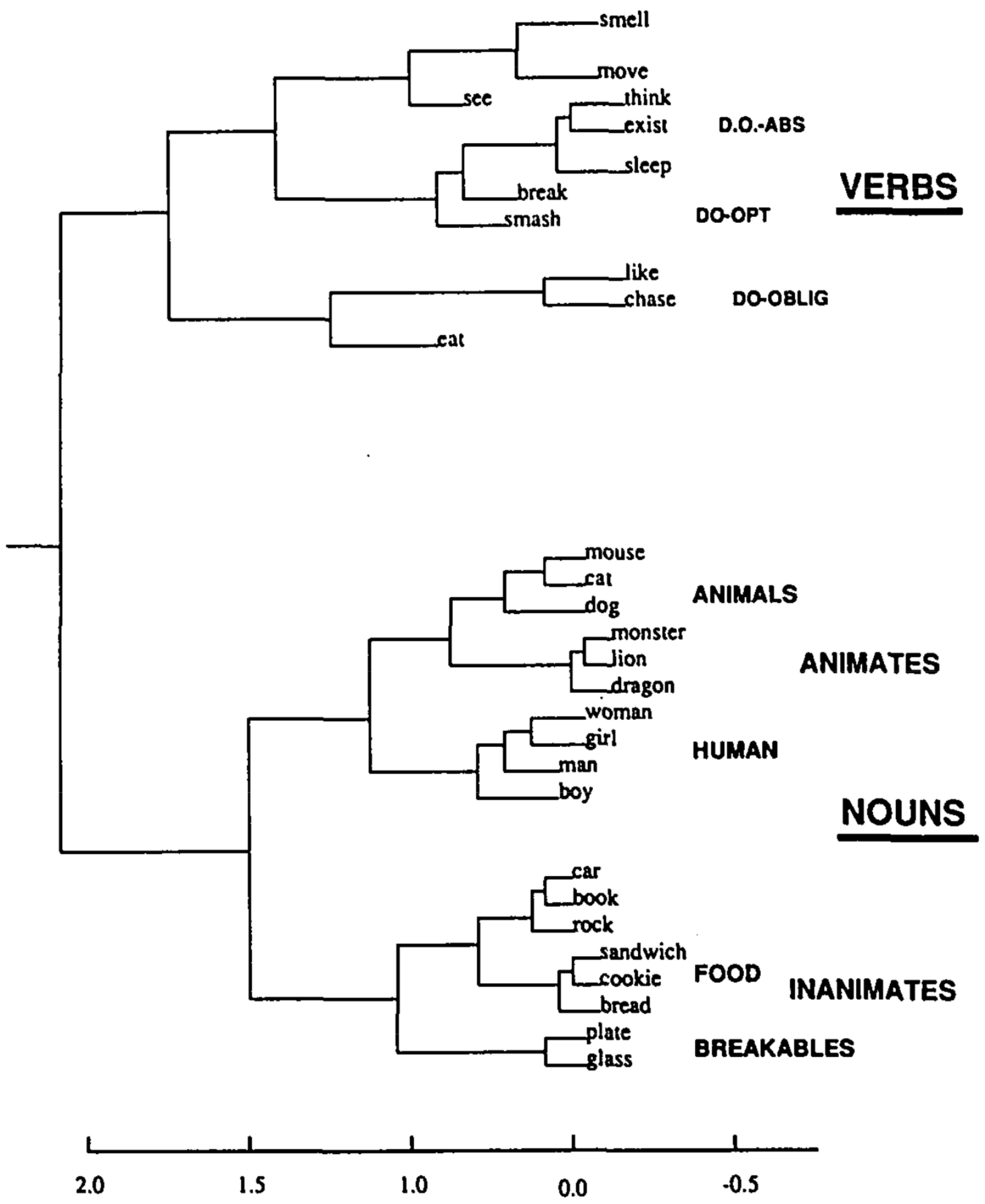


What trends do you notice?

[Elman 1990]

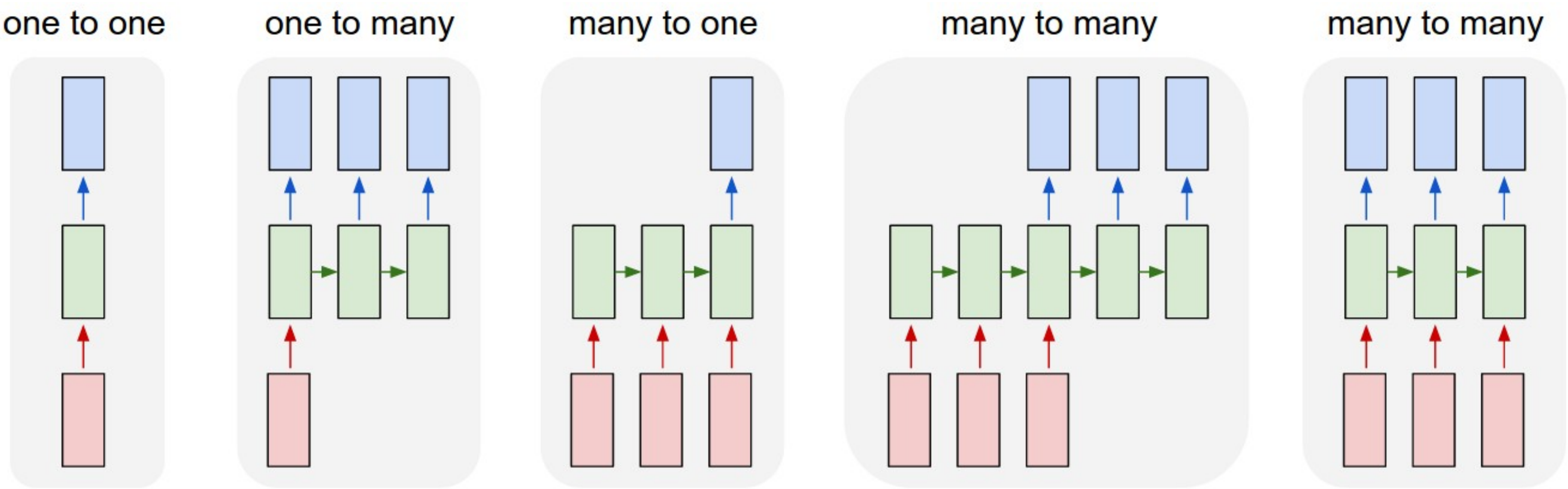
Power of RNNs

Hierarchical clustering of Vanilla RNN hidden states trained as LM on synthetic data:



Elman 1990

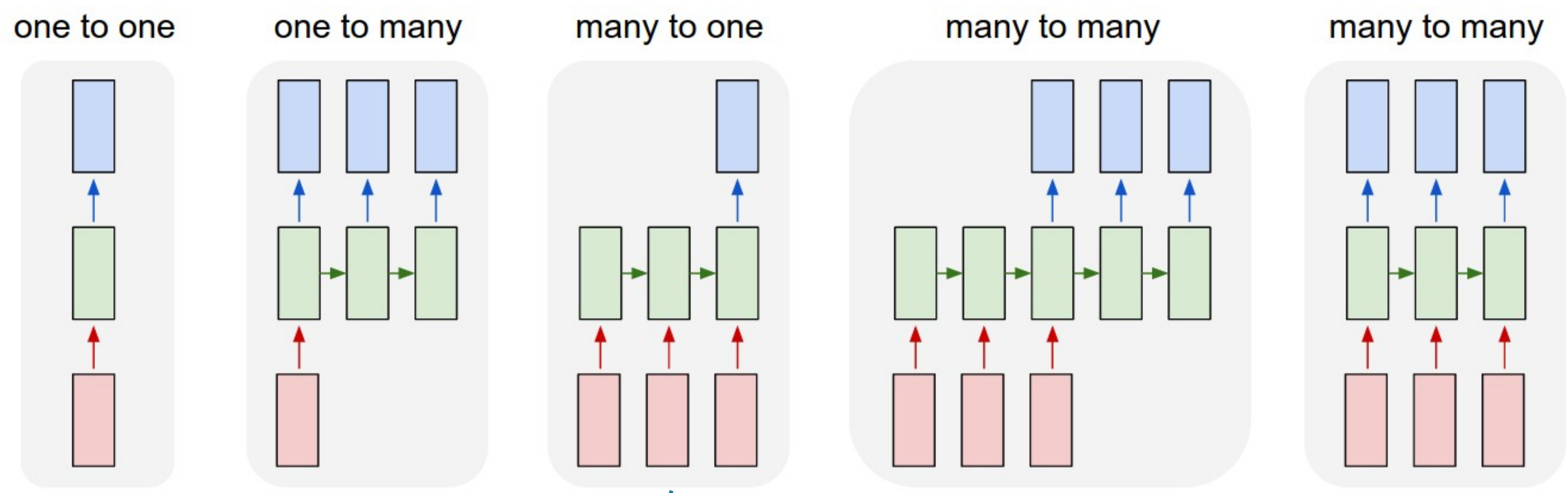
Using RNNs



MLP

e.g. image captioning

Using RNNs

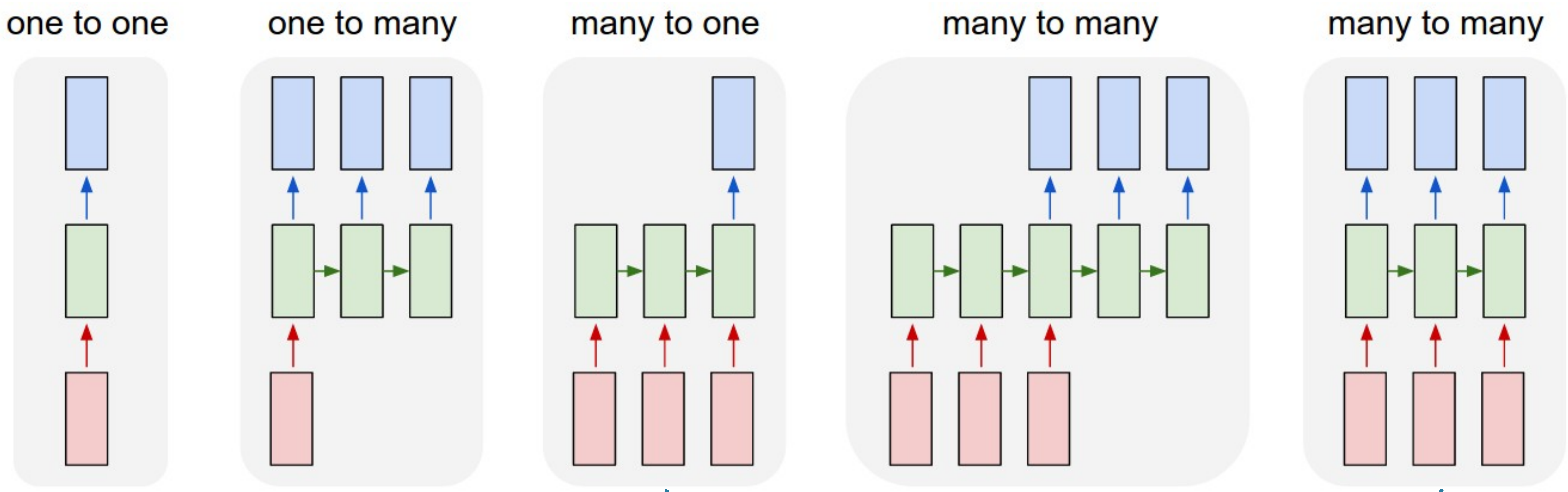


MLP

e.g. image captioning

e.g. text classification

Using RNNs



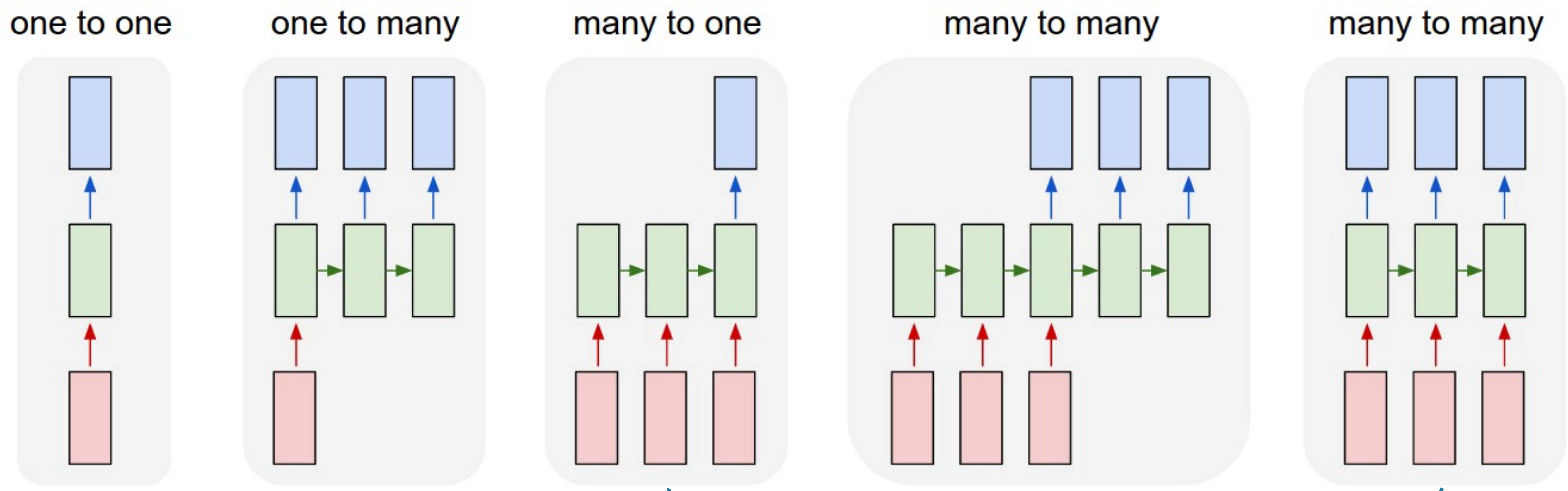
MLP

e.g. image captioning

e.g. text classification

e.g. POS tagging

Using RNNs



MLP

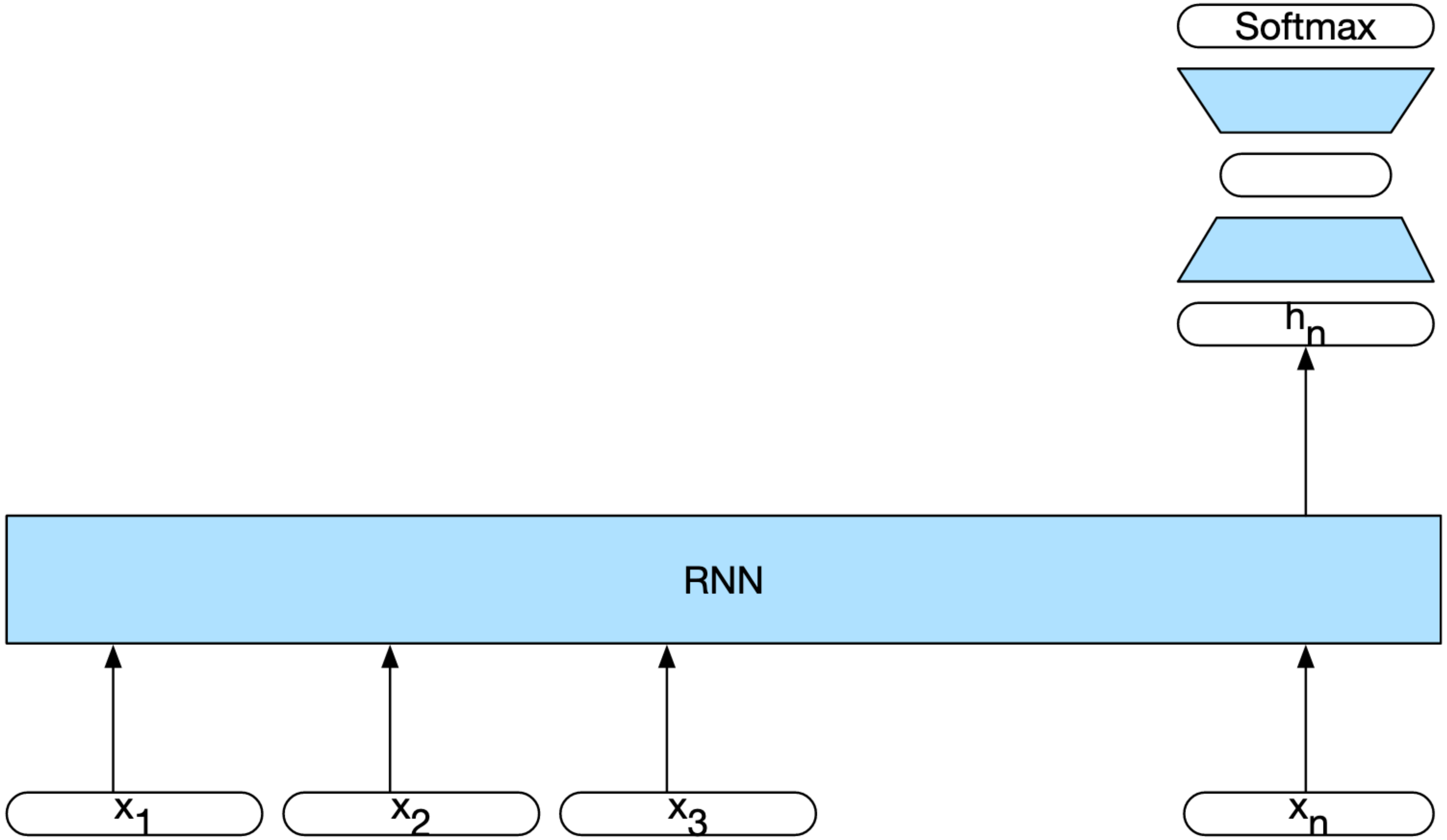
e.g. image captioning

seq2seq (later)

e.g. text classification

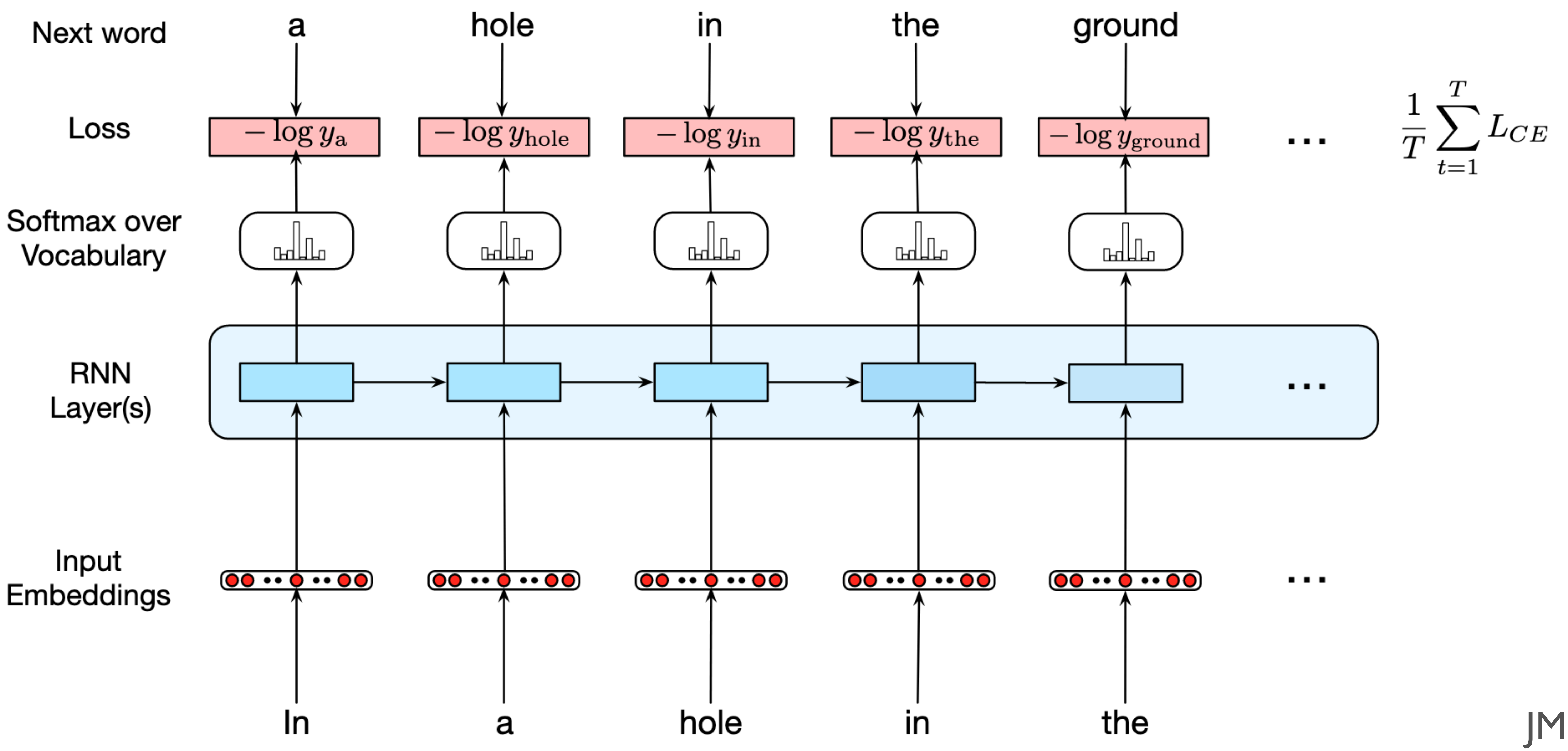
e.g. POS tagging

RNN for Text Classification



JM sec 9.2.5

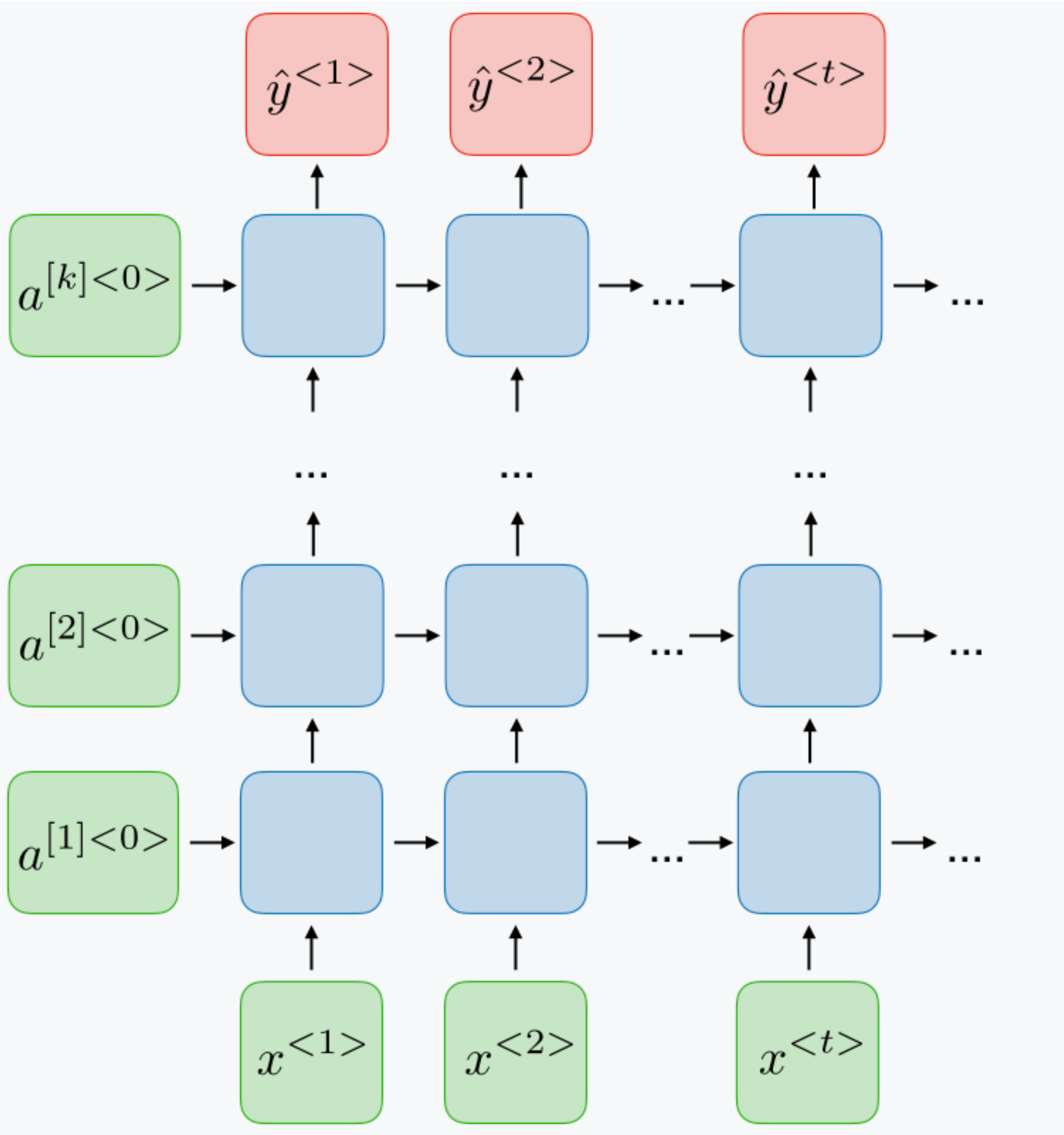
RNNs for Language Modeling



JM sec 9.2.3

Two Extensions

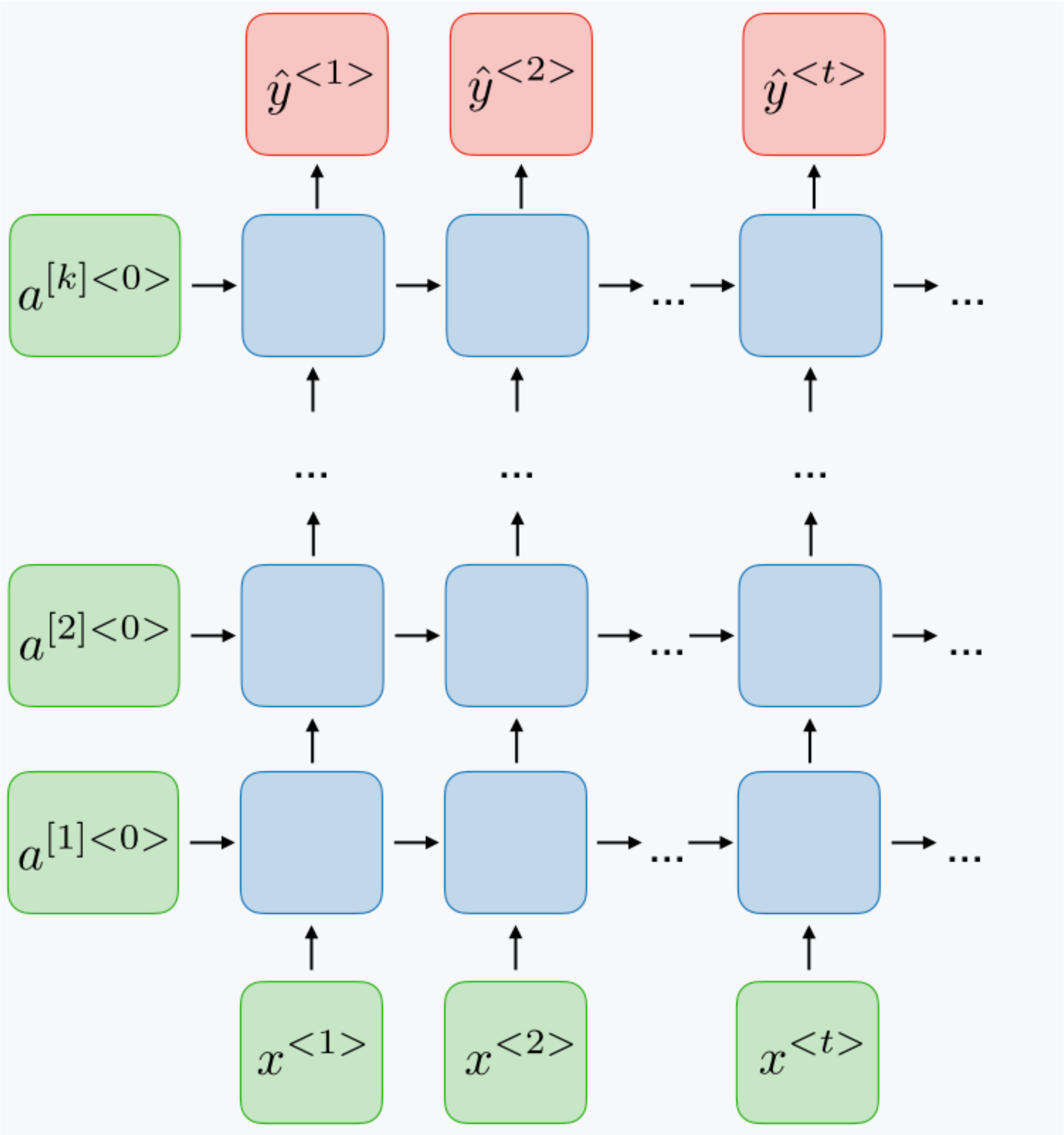
- Deep RNNs:



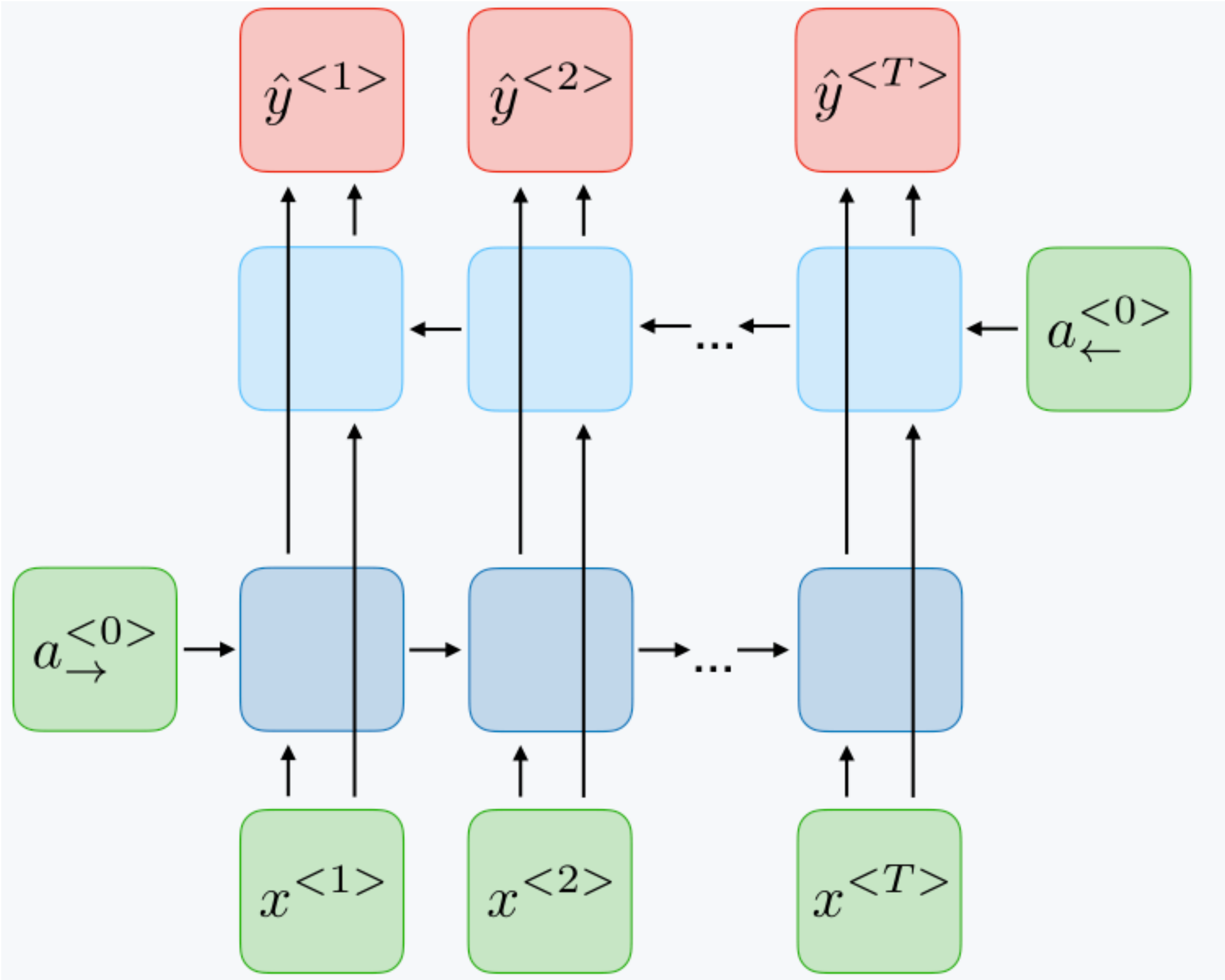
Source: RNN cheat sheet

Two Extensions

- Deep RNNs:



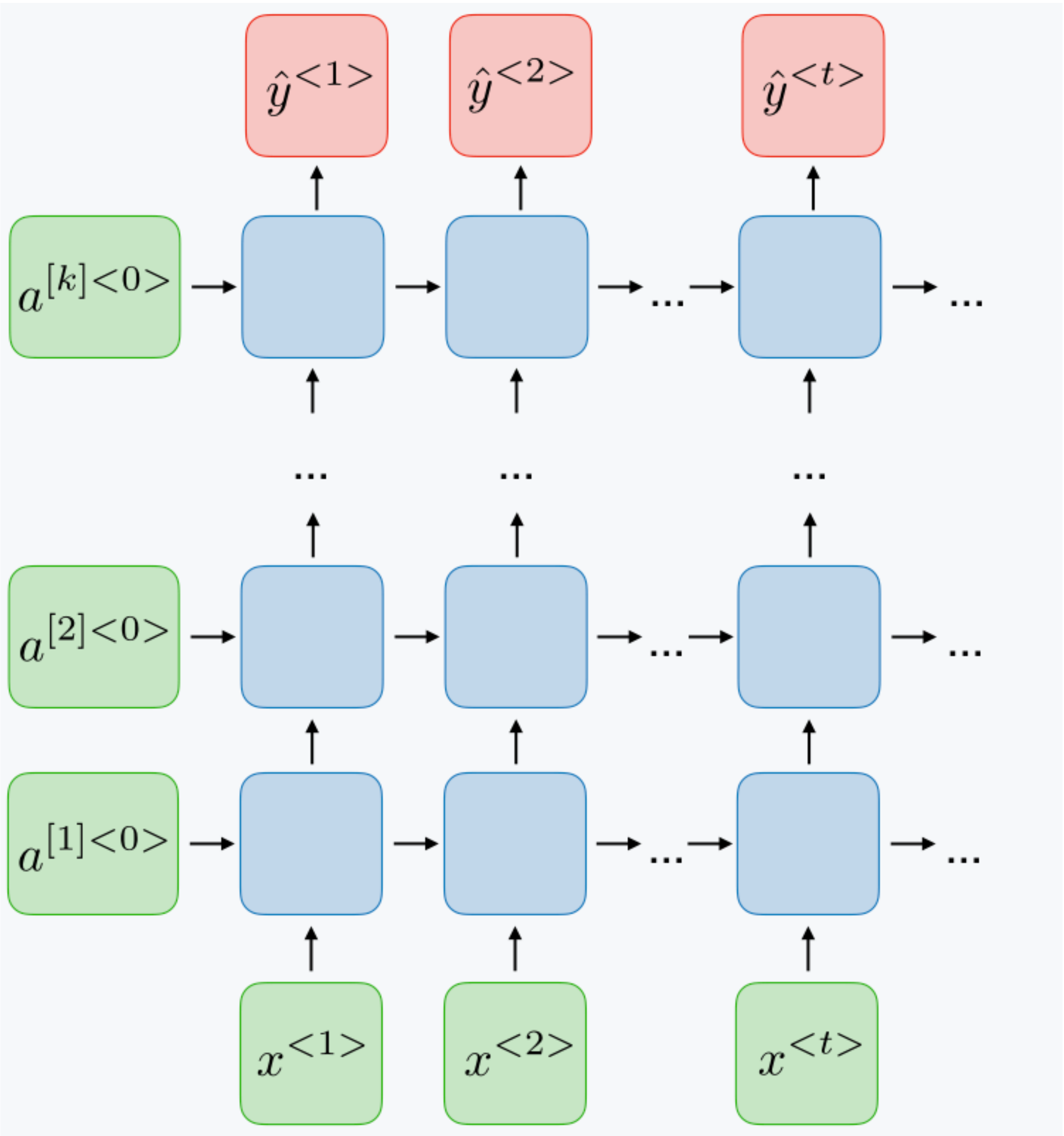
- Bidirectional RNNs:



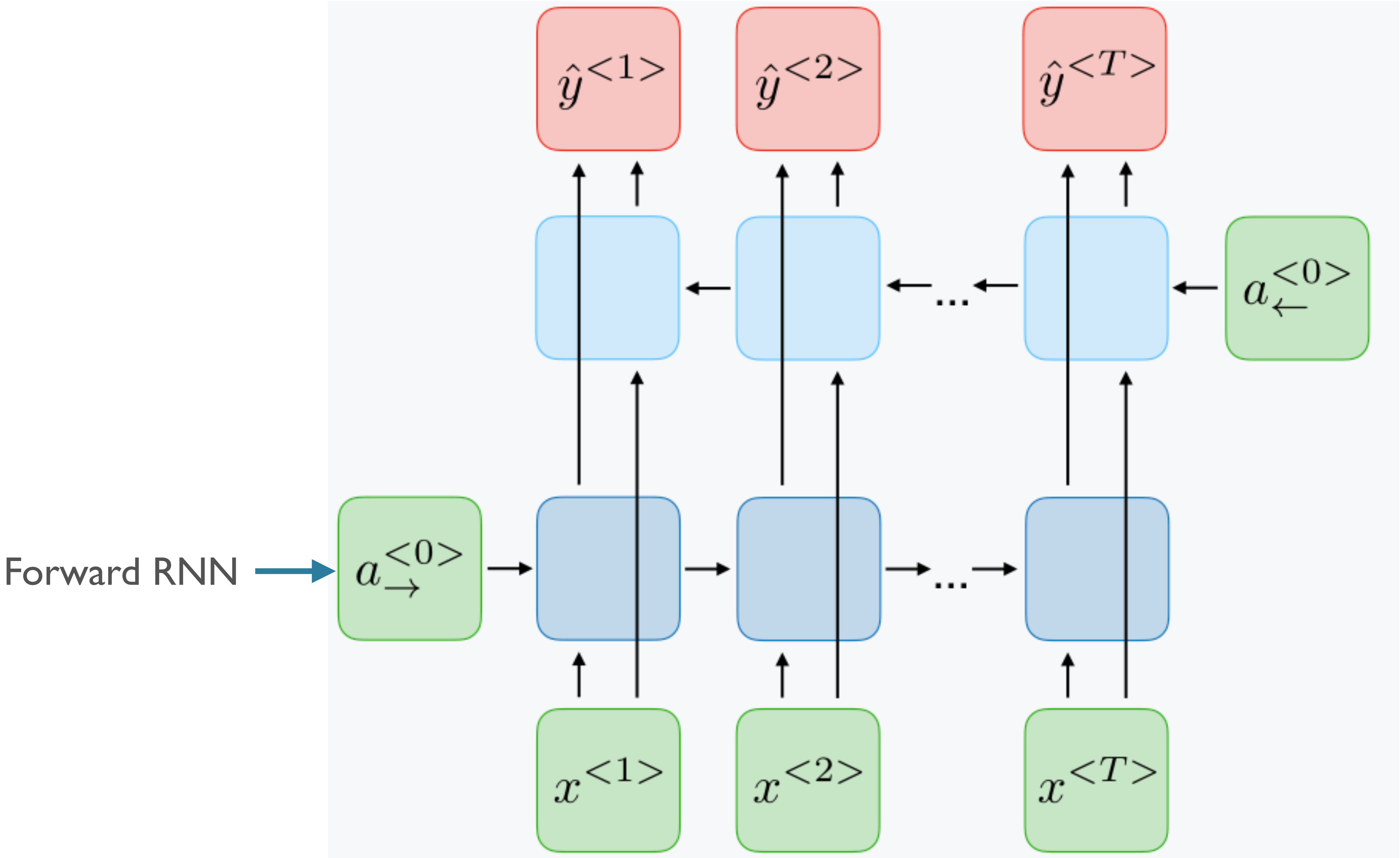
Source: RNN cheat sheet

Two Extensions

- Deep RNNs:



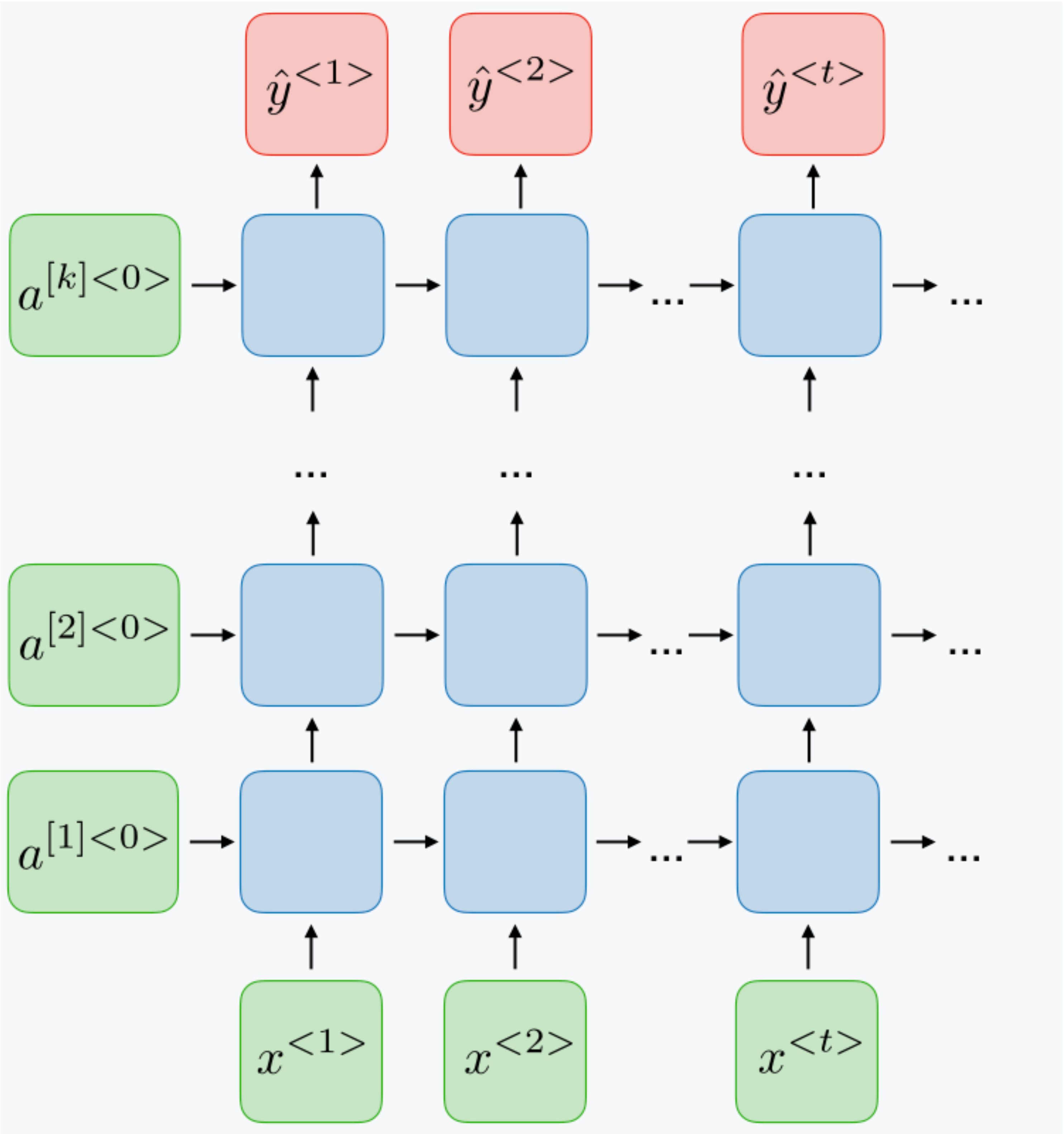
- Bidirectional RNNs:



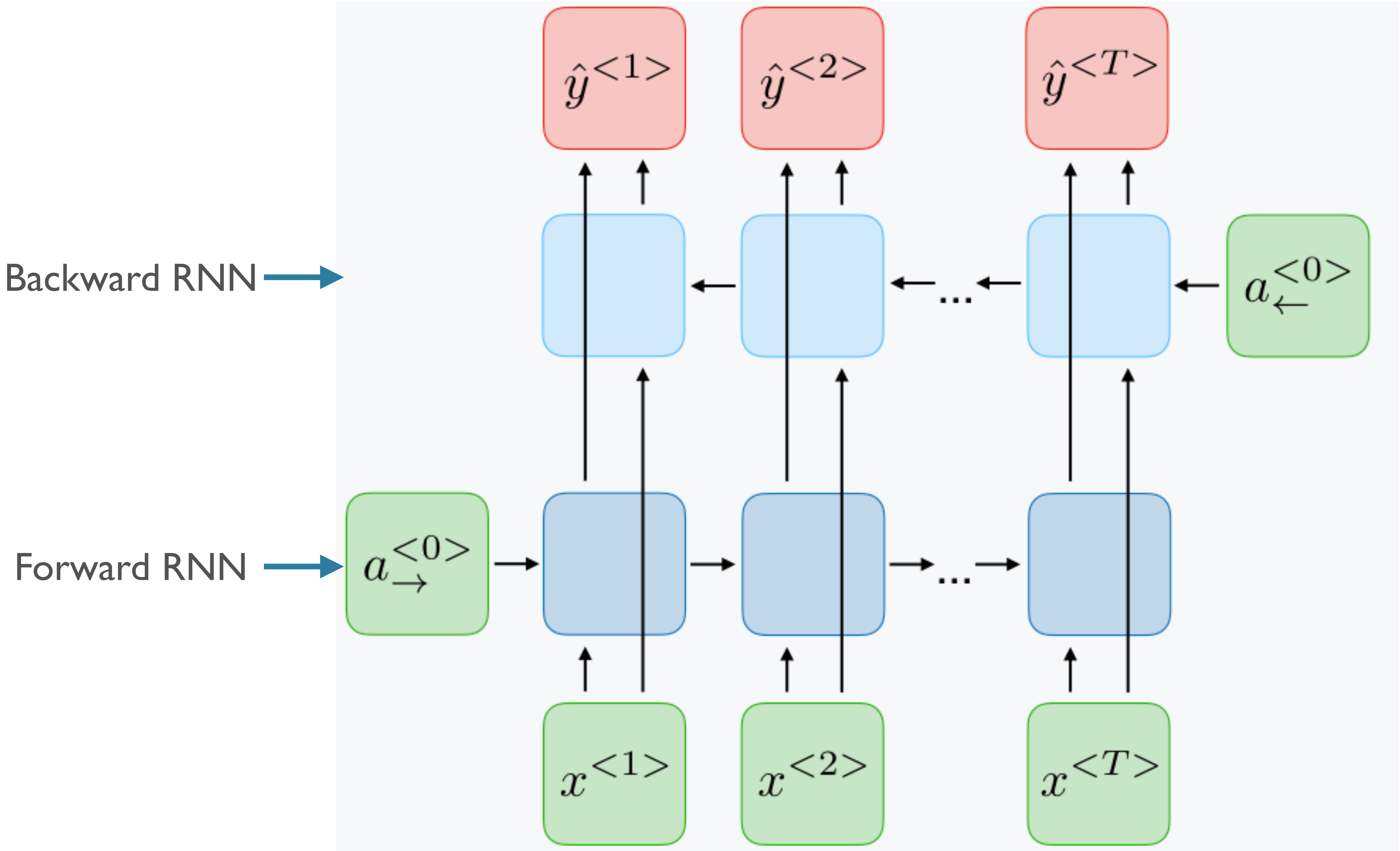
[Source: RNN cheat sheet]

Two Extensions

- Deep RNNs:



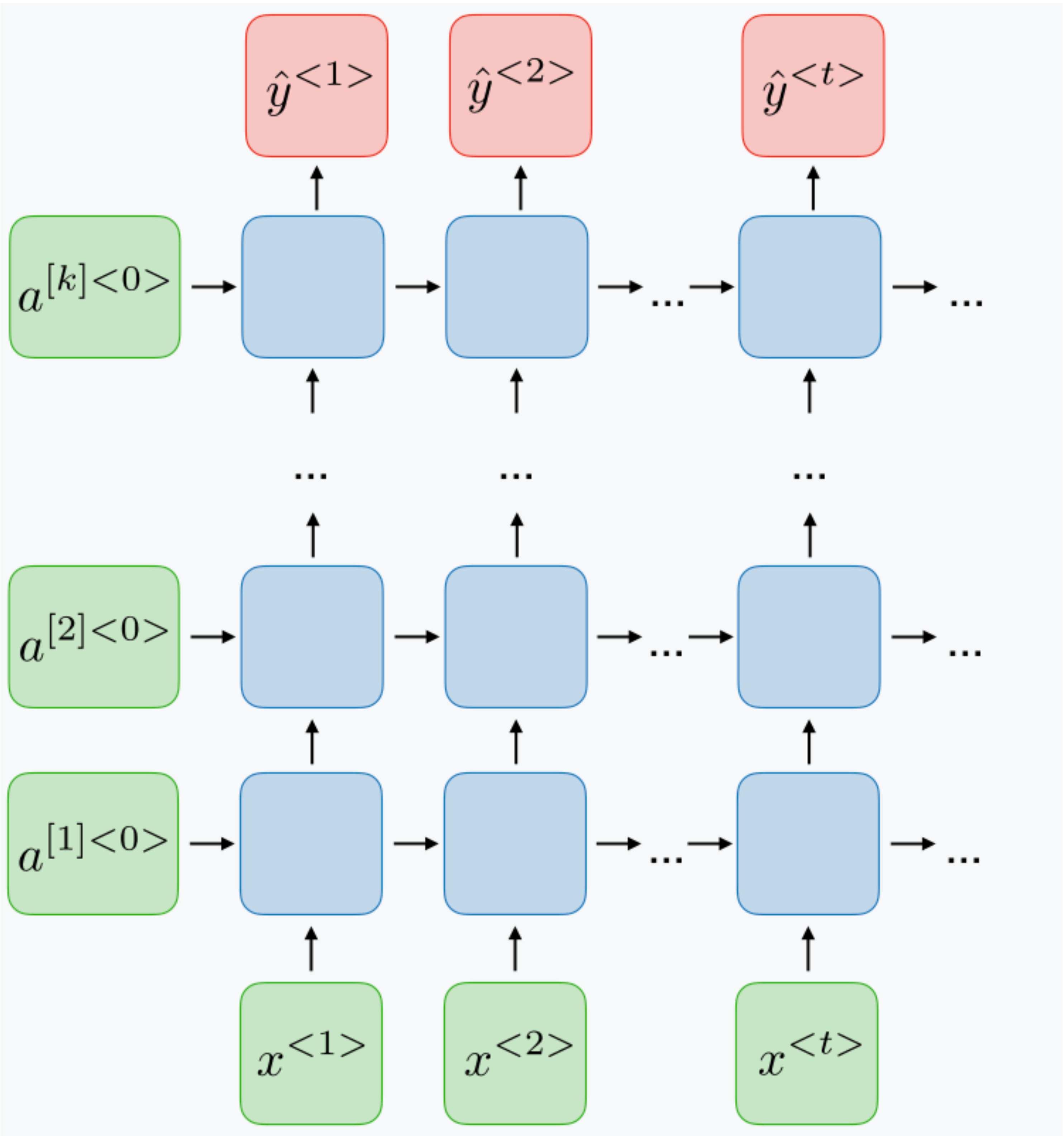
- Bidirectional RNNs:



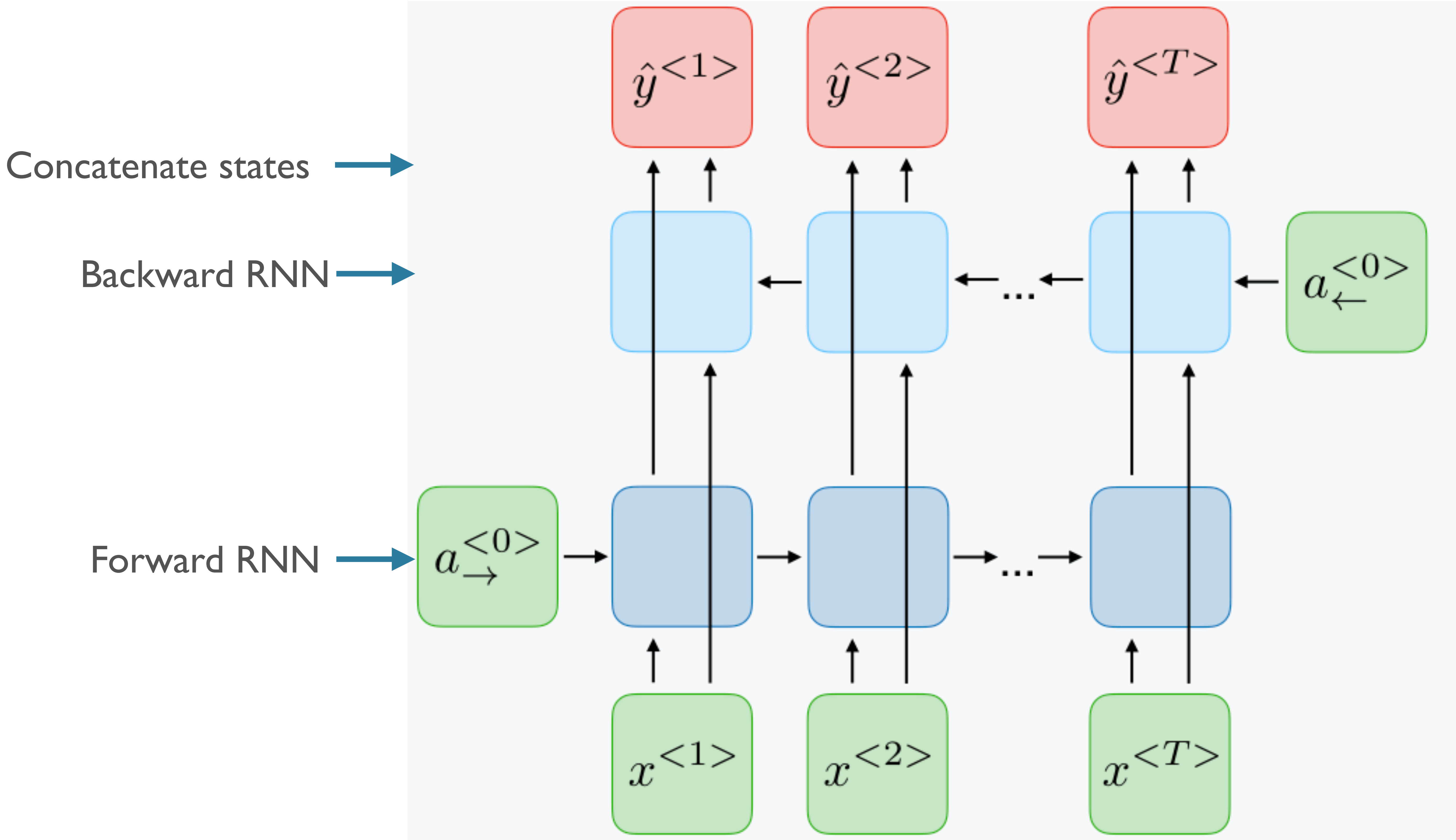
Source: RNN cheat sheet

Two Extensions

- Deep RNNs:



- Bidirectional RNNs:



Source: RNN cheat sheet

Batching in RNNs

- Intuitively, shape of inputs: [batch_size, seq_len, vocab_size]
- But what is sequence length??
  - “This is the first example </s>”: 6
  - “This is another </s>”: 4

Padding and Masking

- Step 1: *pad* all sequences in batch to be of the same length
  - “This is the first example </s>”: 6
  - “This is another </s> PAD PAD”: 6
- Step 2: build a “mask” (1 = True token, 0 = padding)

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

- Step 3: use mask to tell model what to ignore, either
  - Select correct final states [classification]
  - Multiply losses in tagging tasks [LM]

Summary

- RNNs allow for neural processing of *sequential data*
- In principle, should help models capture long-distance dependencies (e.g. number agreement, selectional preferences, …)
  - Maintain a state over time
  - Repeatedly apply the same weights
  - as opposed to n-gram models, which cannot build such dependencies
- Uses: classification, tagging
- Extensions: deep, bidirectional

Next Time

- Discuss a technical problem in training Vanilla RNNs
  - Vanishing gradients
- Introduce *gating-based* RNNs
  - LSTMs
  - GRUs
  - Strengths, weaknesses, differences