Recurrent Neural Networks II

Yao-Yi Chiang
Computer Science and Engineering
University of Minnesota
yaoyi@umn.edu

Recall: Long-Short Term Memory

- LSTM uses a complex recurrent unit with gates to control what information is passed through that can avoid vanishing gradient problem in standard RNNs

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$$

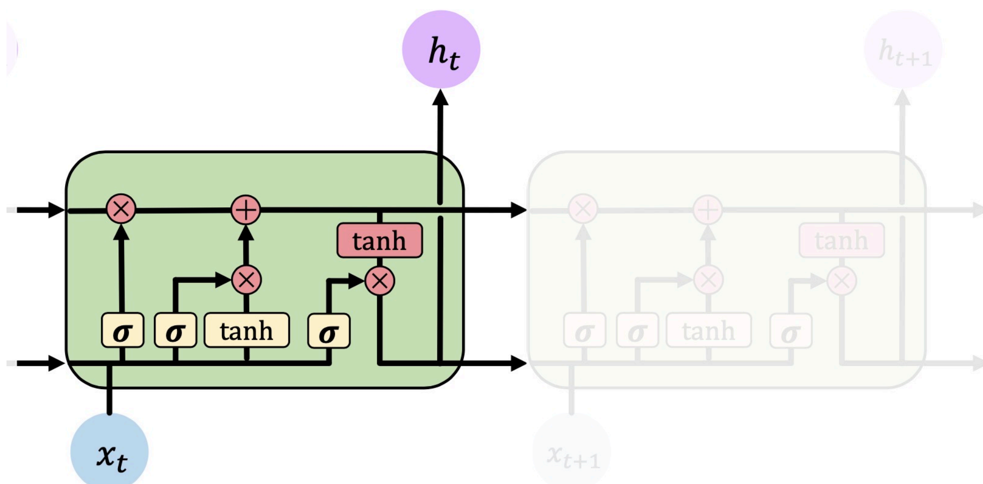
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$



Gated Recurrent Unit (GRU)

- GRU is another RNN variant [Cho et al. 2014]
  - GRU reduces the number of gates but achieves similar performance to LSTM

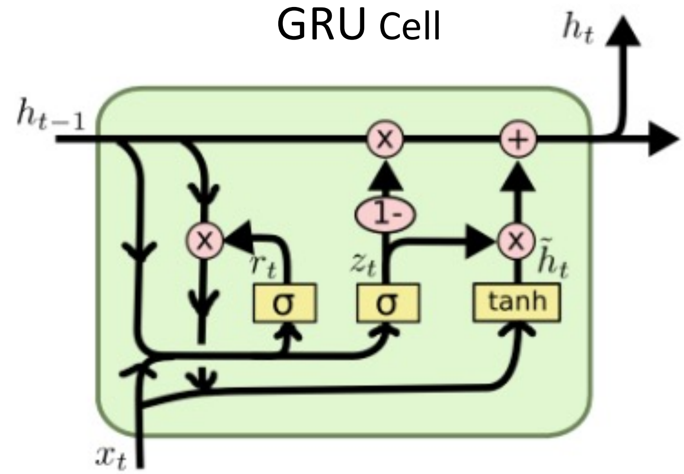
$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r)$$

$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h)$$

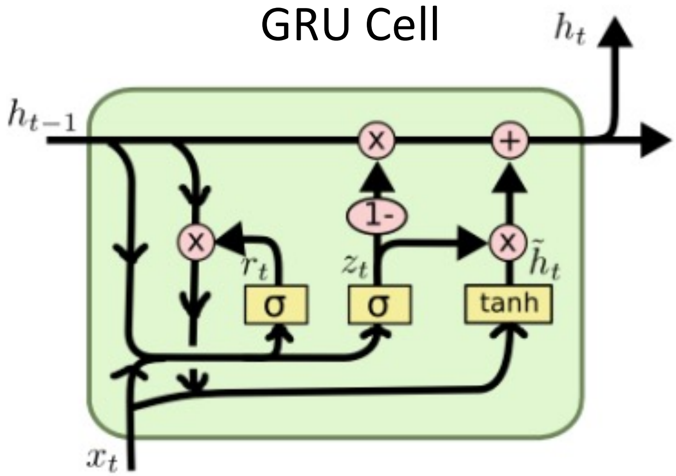
$$h_t = \mathrm{z}_t \odot h_{t-1} + (1 - \mathrm{z}_t) * \tilde{h}_t$$

GRU Cell



GRU: Update Gate

GRU Cell

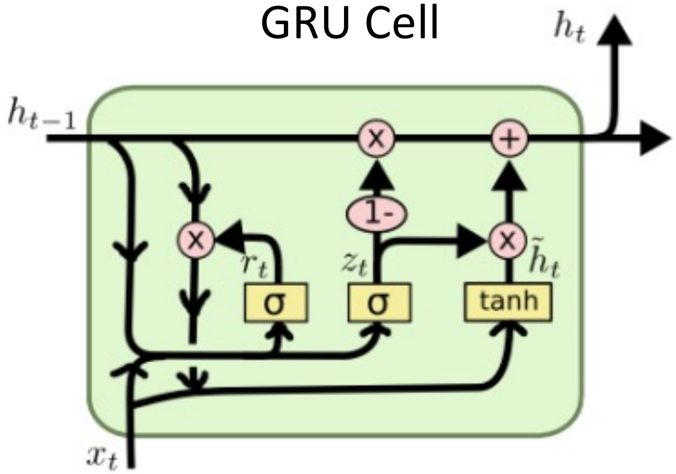


$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z)$$

- Concatenate previous hidden state and current input
- Update gate controls what parts of hidden state are updated (used as z_t) VS. preserved (used as $(1 - z_t)$)

GRU: Reset Gate

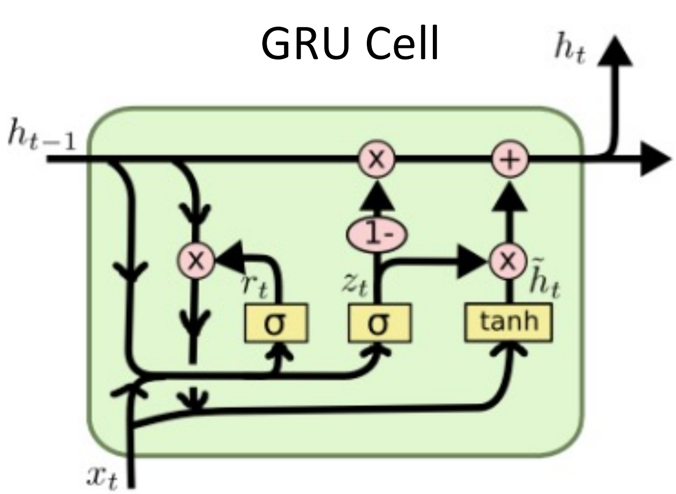
GRU Cell



$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r)$$

- Reset gate controls what parts of previous hidden state are used to compute new content

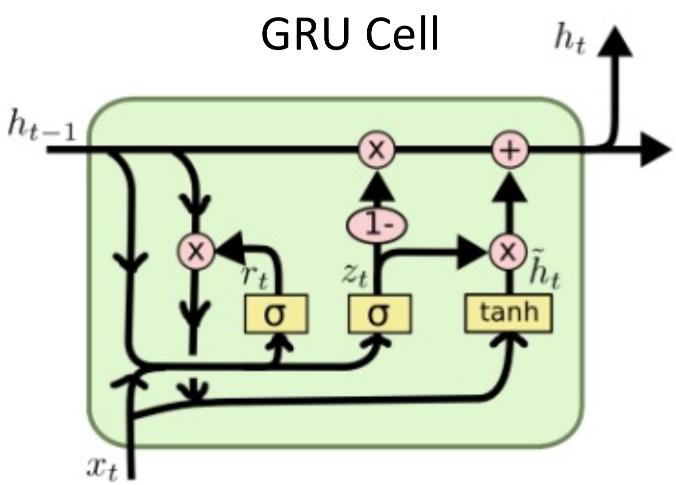
GRU: New Hidden State Content



$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h)$$

- r_t selects useful parts of previous hidden state
- Use $r_t \odot h_{t-1}$ and current input to compute new hidden content

GRU: Output Hidden State



$$h_t = z_t \odot h_{t-1} + (1 - z_t) * \tilde{h}_t$$

- Update gate simultaneously controls what is kept from previous hidden state, and what is updated to new hidden state content

GRU VS. LSTM: Which to Use?

- In many tasks, both architectures yield comparable performance [1]
- Both architectures were proposed to tackle the vanishing gradient problem but using a different way of **fusing previous timestep information with gates** to prevent from vanishing gradients
- Nevertheless, the gradient flow in LSTM comes from three different gates, so intuitively, you would observe **more variability in the gradient descent** compared to GRUs

[1] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222–2232.

GRU VS. LSTM: Which to Use?

- GRU has two gates (reset and update gates) whereas an LSTM has three gates (namely input, output, and forget gates)
- **GRU is considered more efficient** in terms of simpler structure with fewer parameters
- In small-scale datasets with not too long sequences, it is common to use GRU cells since with fewer data the expressive power of LSTM may not be exposed

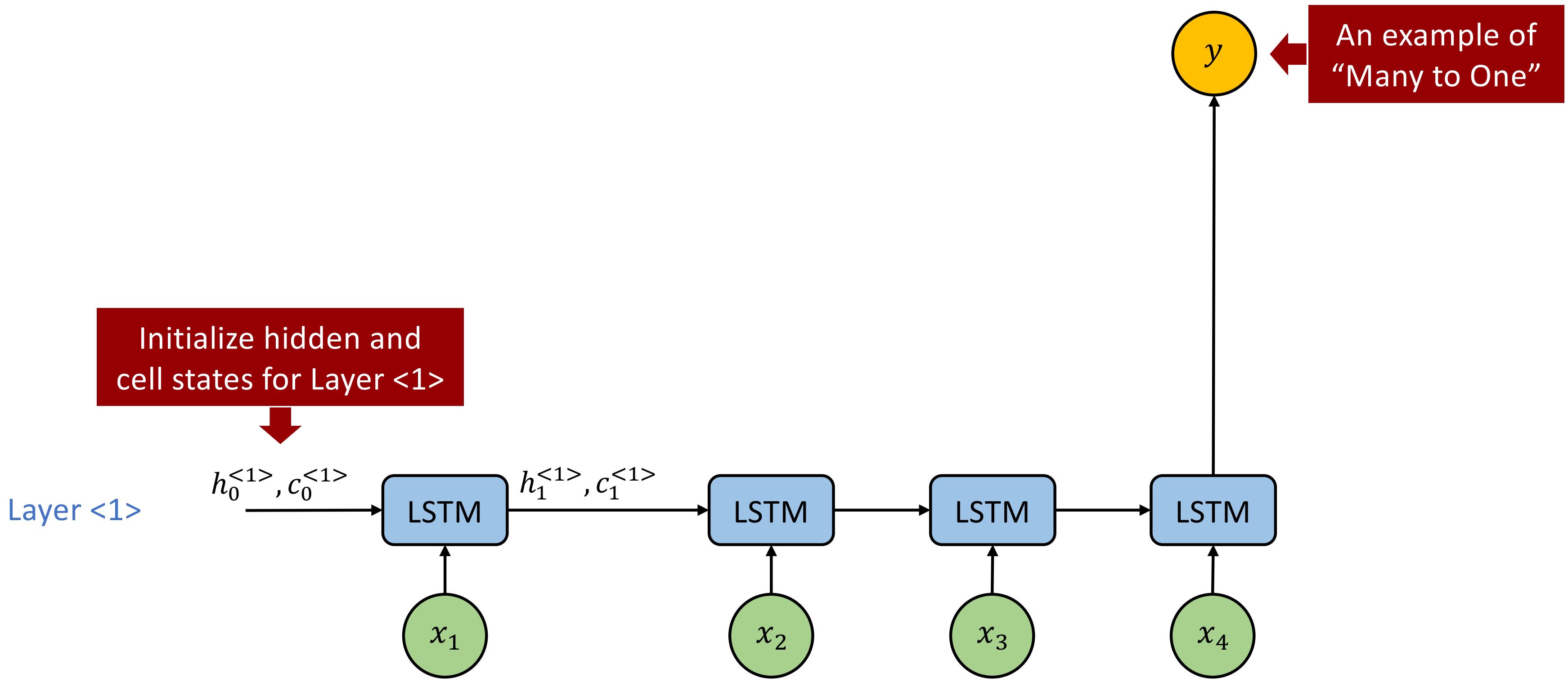
GRU VS. LSTM: Which to Use?

- If you deal with large datasets, the greater expressive power of LSTMs may lead to superior results

- In theory, the LSTM cells should **remember longer sequences** than GRUs and outperform them in tasks requiring modeling long-range correlations

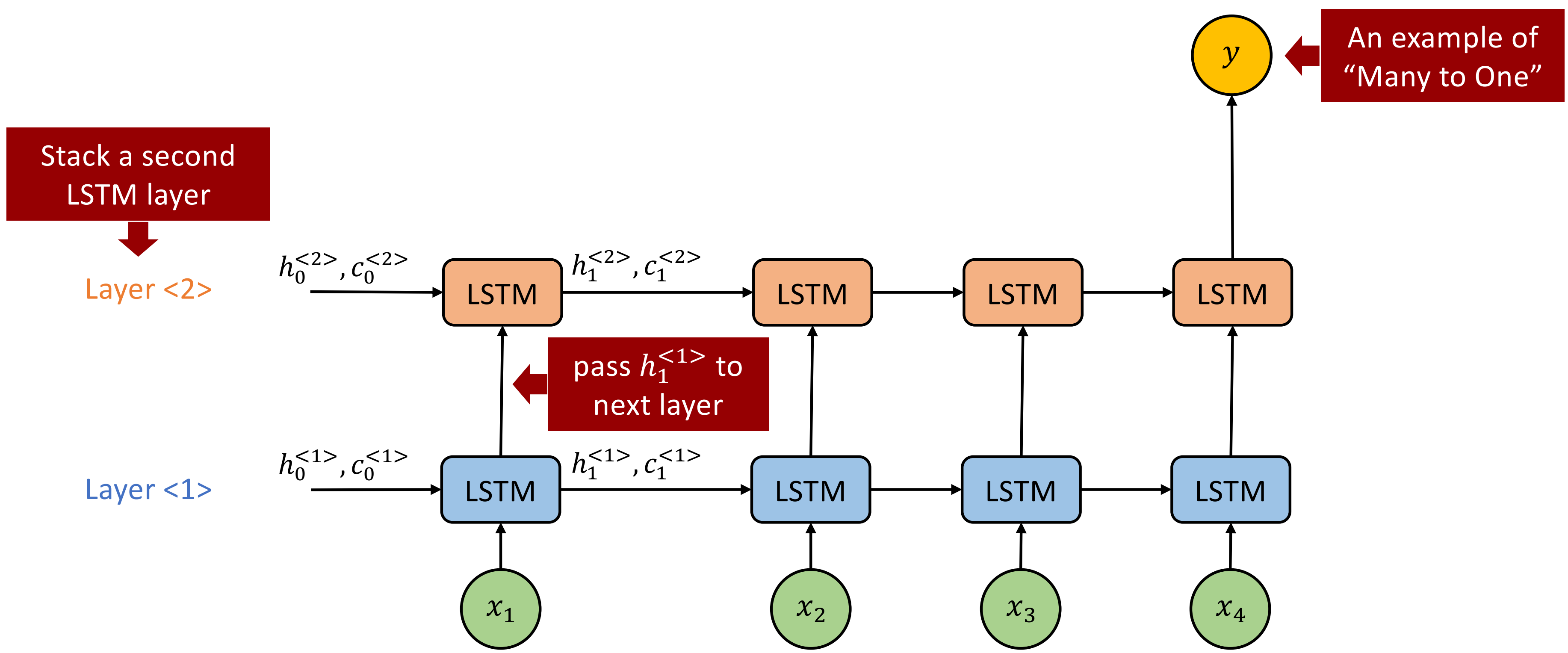
- Which to use **lies in the data**

Deep RNN Examples: Stacked LSTMs



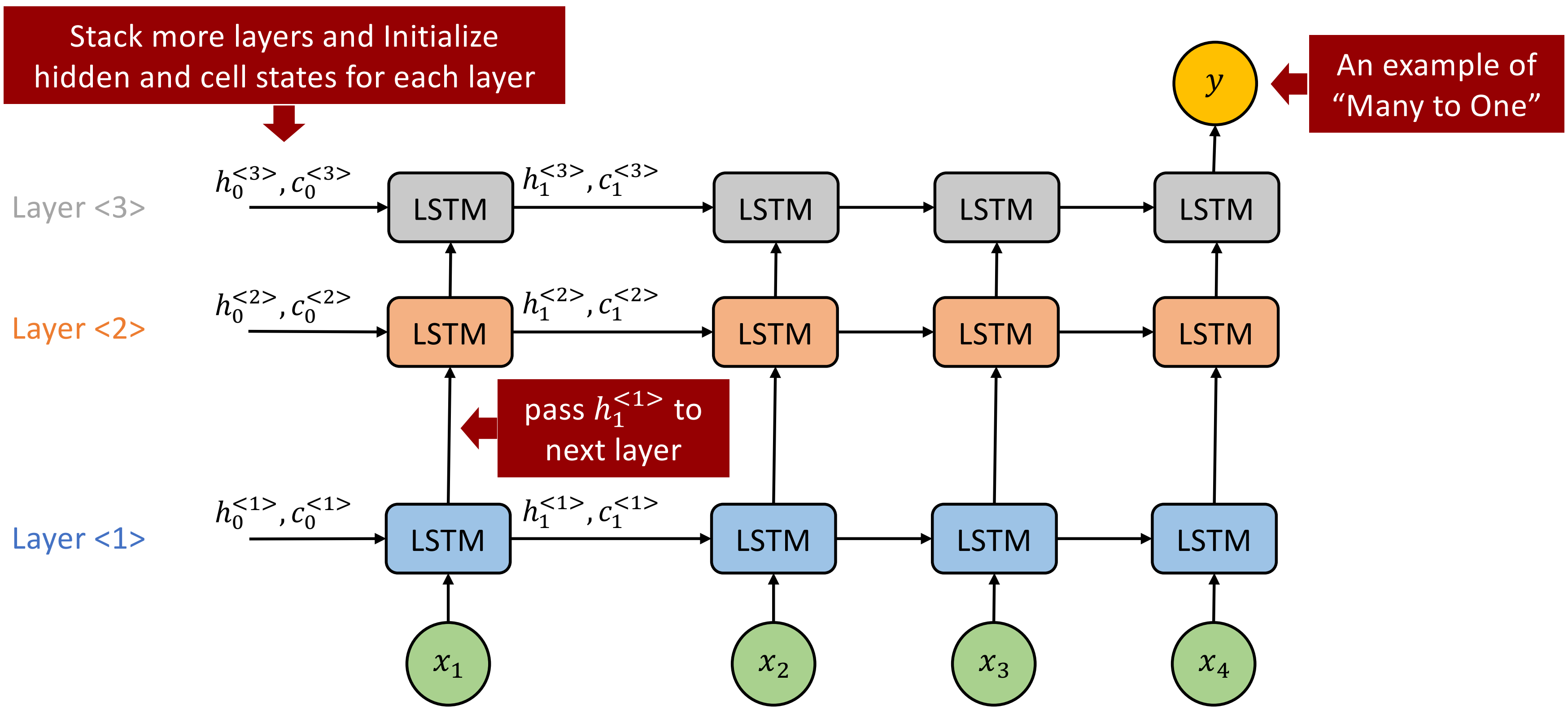
* LSTM cell can be replaced with other RNN cells

Deep RNN Examples: Stacked LSTMs



* LSTM cell can be replaced with other RNN cells

Deep RNN Examples: Stacked LSTMs



* LSTM cell can be replaced with other RNN cells

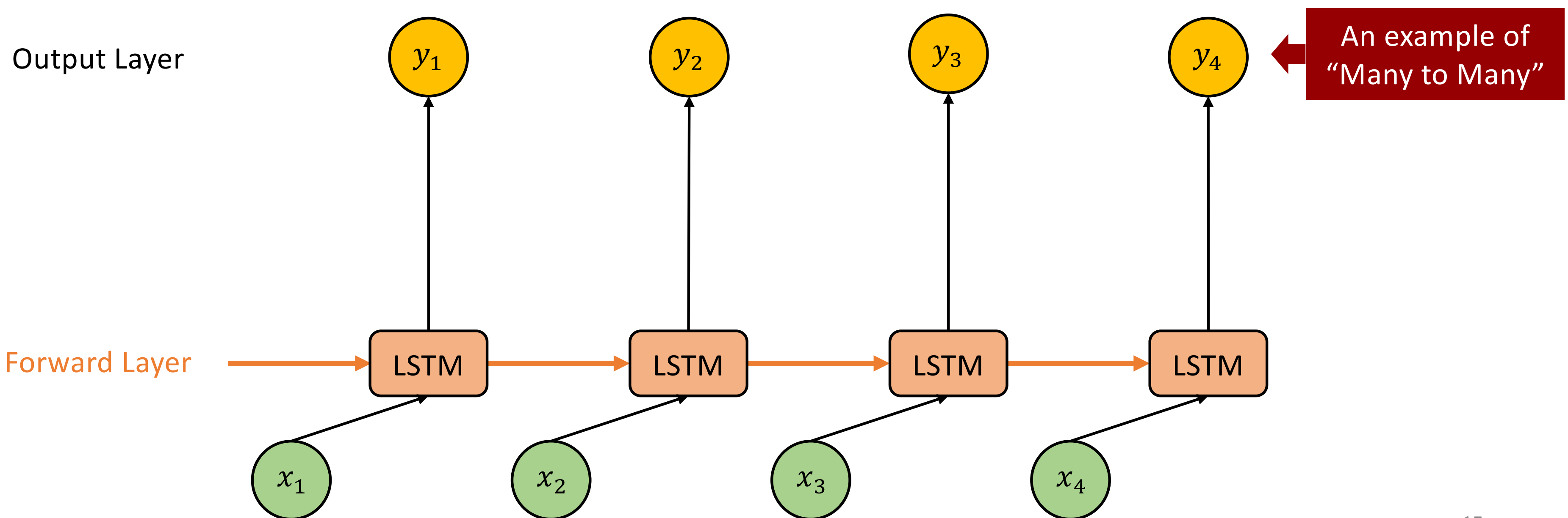
Stacked LSTMs

- Staked LSTMs were first introduced in [1] for speech recognition
  - They also found that the depth of the network was more important than the number of memory cells in a layer to model skill
- Why increasing depth?
  - Given that Stacked LSTMs operate on sequence data, the addition of layers adds levels of abstraction of input observations over time
- Other domains
  - Traffic forecast [2], weather forecast [3]

[1] Graves, A., Mohamed, A. R., & Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649).
[2] Du, X., Zhang, H., Van Nguyen, H., & Han, Z. Stacked LSTM deep learning model for traffic prediction in vehicle-to-vehicle communication. In 2017 IEEE 86th Vehicular Technology Conference (pp. 1-5).
[3] Karevan, Z., & Suykens, J. A. (2018). Spatio-temporal stacked LSTM for temperature prediction in weather forecasting. arXiv preprint arXiv:1811.06341.

Deep RNN Examples: Bidirectional LSTM

- Regular LSTM considers forward direction, i.e., past to future



Bidirectional LSTM Example

- Name Entity Recognition

person fruit

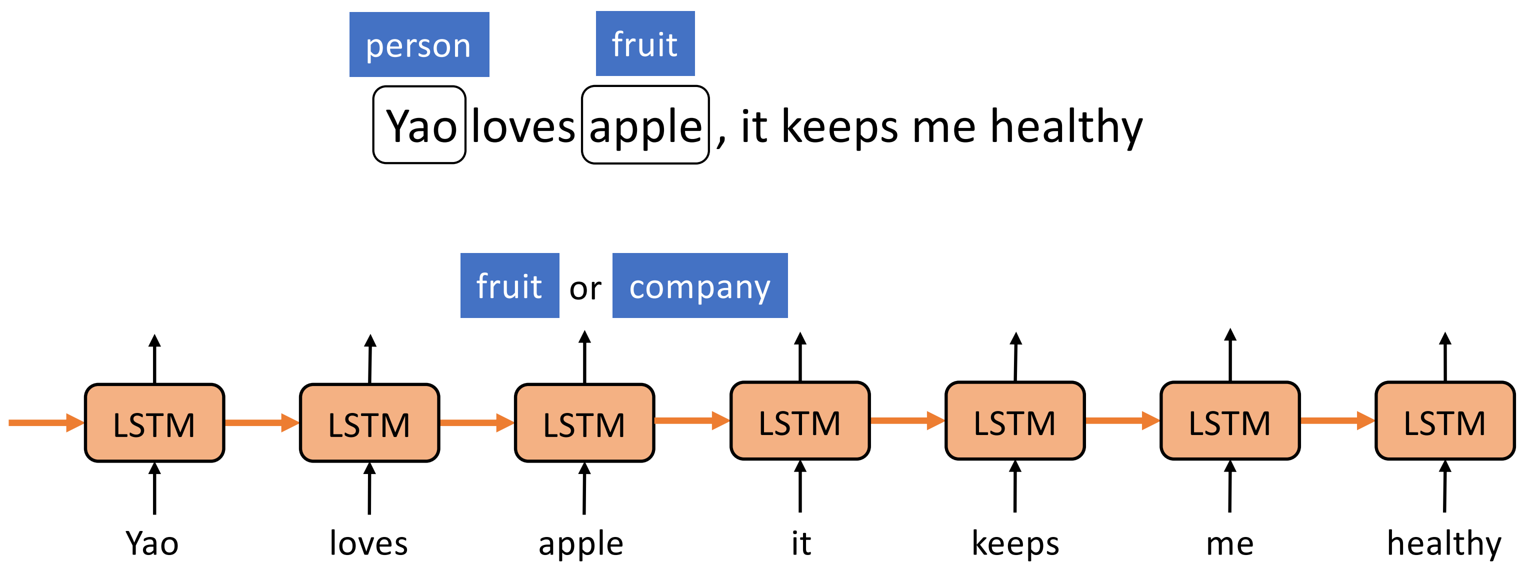
Yao loves apple, it keeps me healthy

person company

Yao loves apple, it produces the best electronics

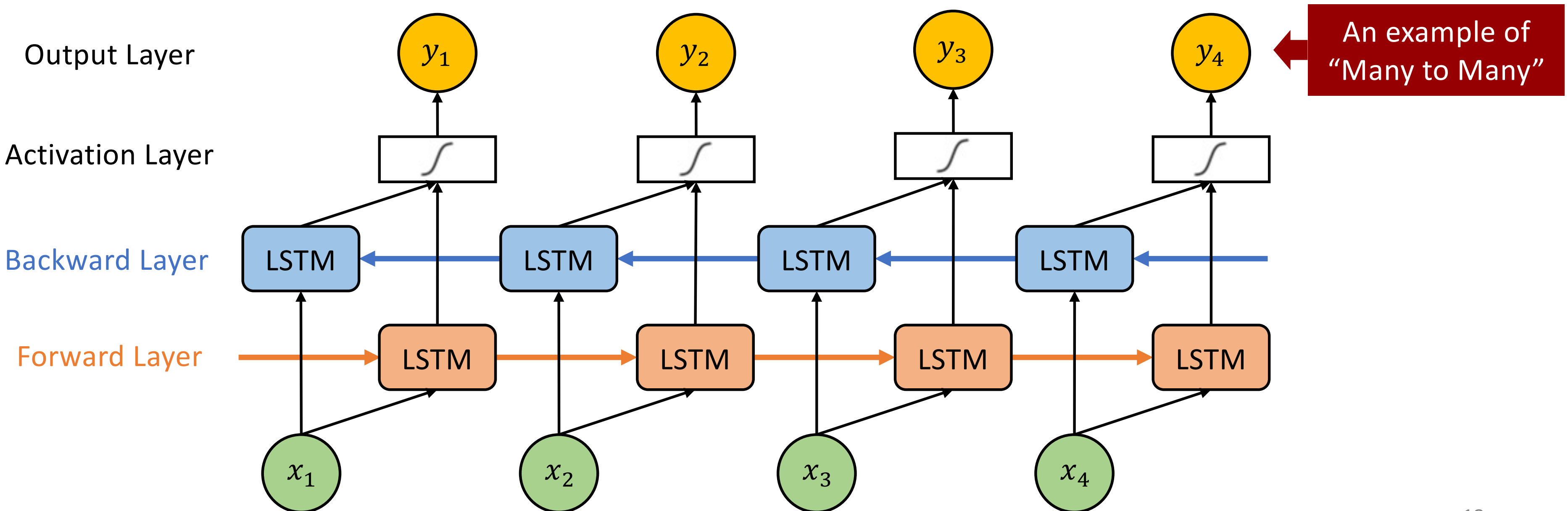
Bidirectional LSTM Example

- Name Entity Recognition
  - Regular LSTM networks might not work



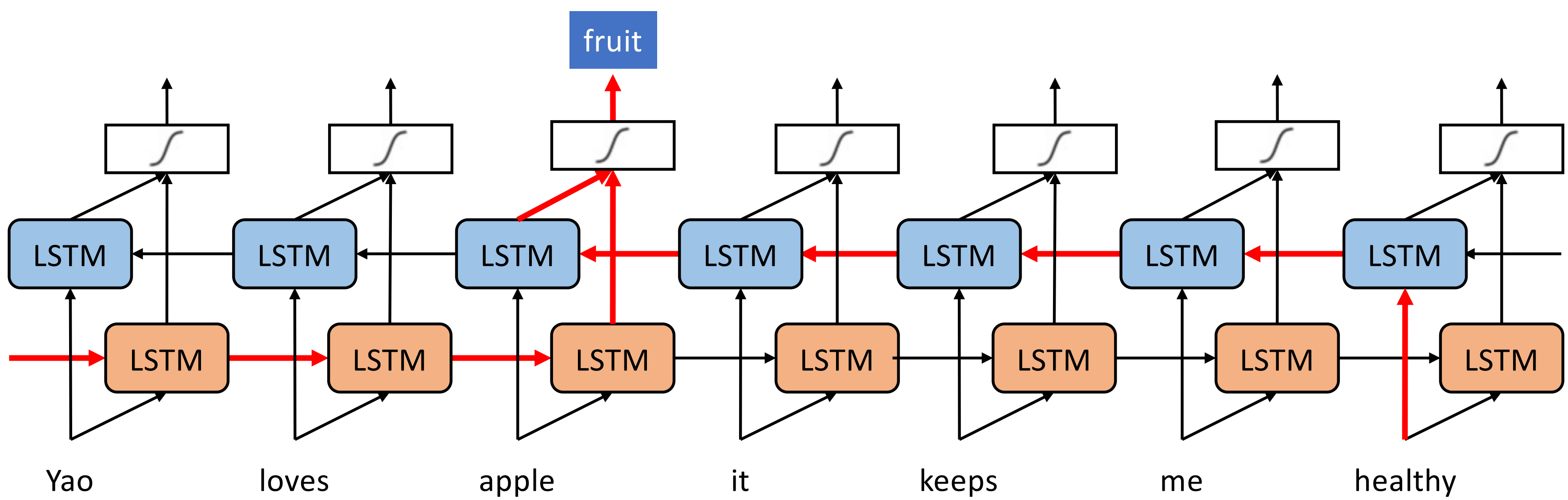
Bidirectional LSTM

- Bidirectional LSTM considers the sequence information in both directions backwards (future to past) and forward (past to future)



Bidirectional LSTM Example

- Name Entity Recognition
  - Red arrows are the information flow



RNN Applications: Sentiment Classification

Ratemyprofessors Sentiment Classification

AWESOME

QUALITY
5.0

DIFFICULTY
4.0

INF553

For Credit: **Yes** Attendance: **Not Mandatory** Would Take Again: **Yes**

Textbook: **Yes** Online Class: **Yes**

Prof Chiang is really a good professor who truly cares his students. Data Mining is a hard course, but he explained every concepts very clear, and there are lots of chance you can get extra credit. If you work hard, you will get good grade (and knowledge).

AWFUL

QUALITY
1.0

DIFFICULTY
5.0

INF553

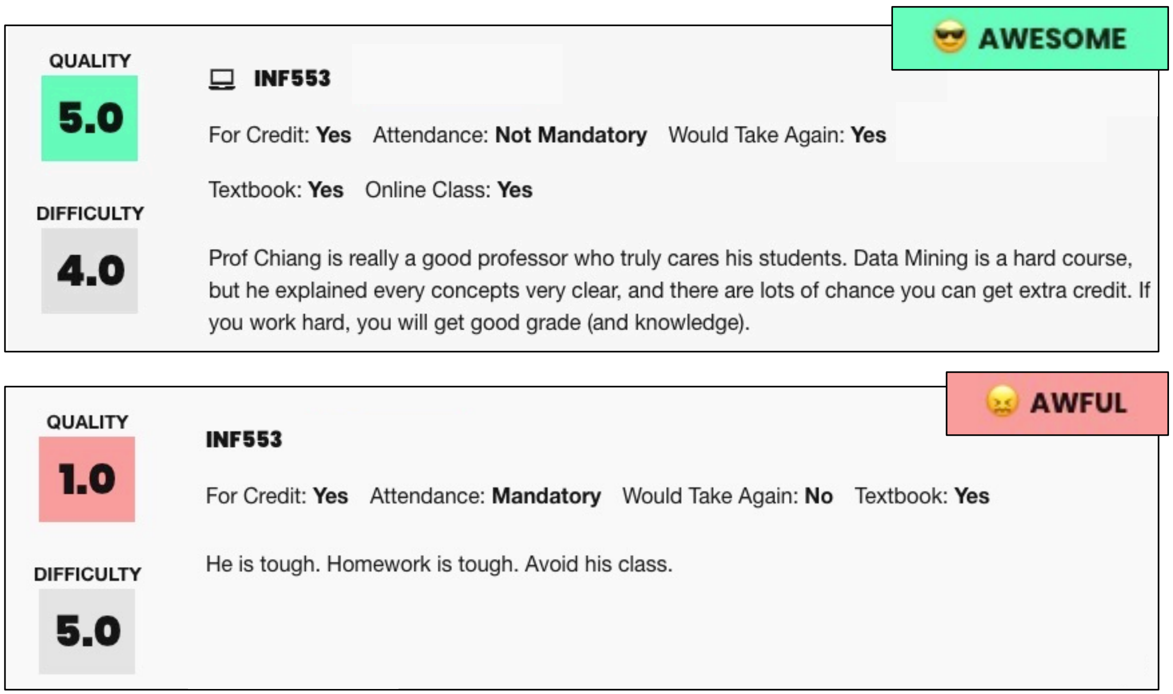
For Credit: **Yes** Attendance: **Mandatory** Would Take Again: **No** Textbook: **Yes**

He is tough. Homework is tough. Avoid his class.

Source: https://www.ratemyprofessors.com/ShowRatings.jsp?tid=2391740

RNN Applications: Sentiment Classification

Ratemyprofessors Sentiment Classification


AWESOME

QUALITY
5.0

DIFFICULTY
4.0

INF553

For Credit: **Yes** Attendance: **Not Mandatory** Would Take Again: **Yes**

Textbook: **Yes** Online Class: **Yes**

Prof Chiang is really a good professor who truly cares his students. Data Mining is a hard course, but he explained every concepts very clear, and there are lots of chance you can get extra credit. If you work hard, you will get good grade (and knowledge).

AWFUL

QUALITY
1.0

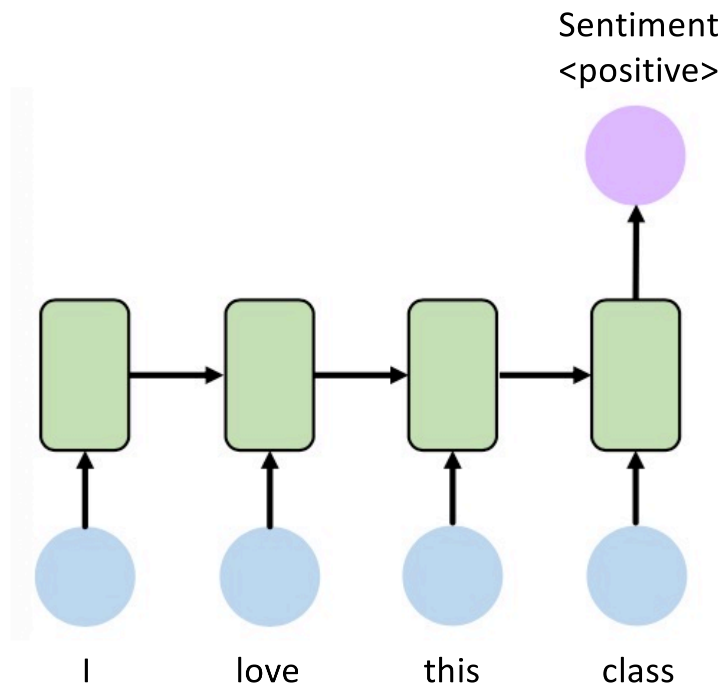
DIFFICULTY
5.0

INF553

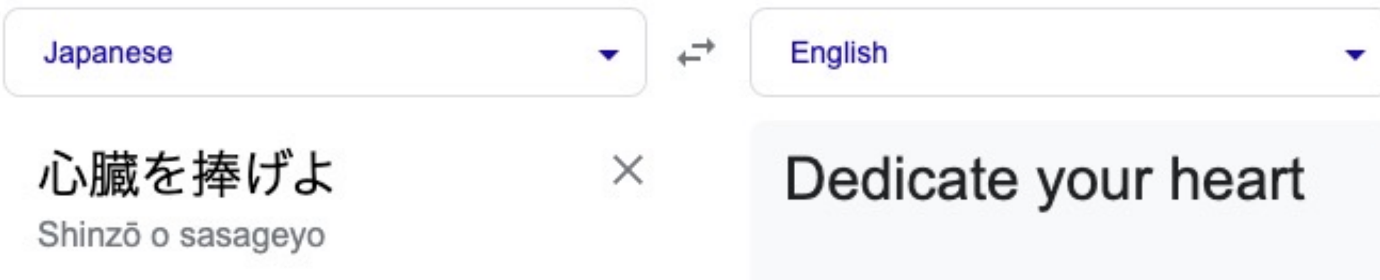
For Credit: **Yes** Attendance: **Mandatory** Would Take Again: **No** Textbook: **Yes**

He is tough. Homework is tough. Avoid his class.

Source: https://www.ratemyprofessors.com/ShowRatings.jsp?tid=2391740

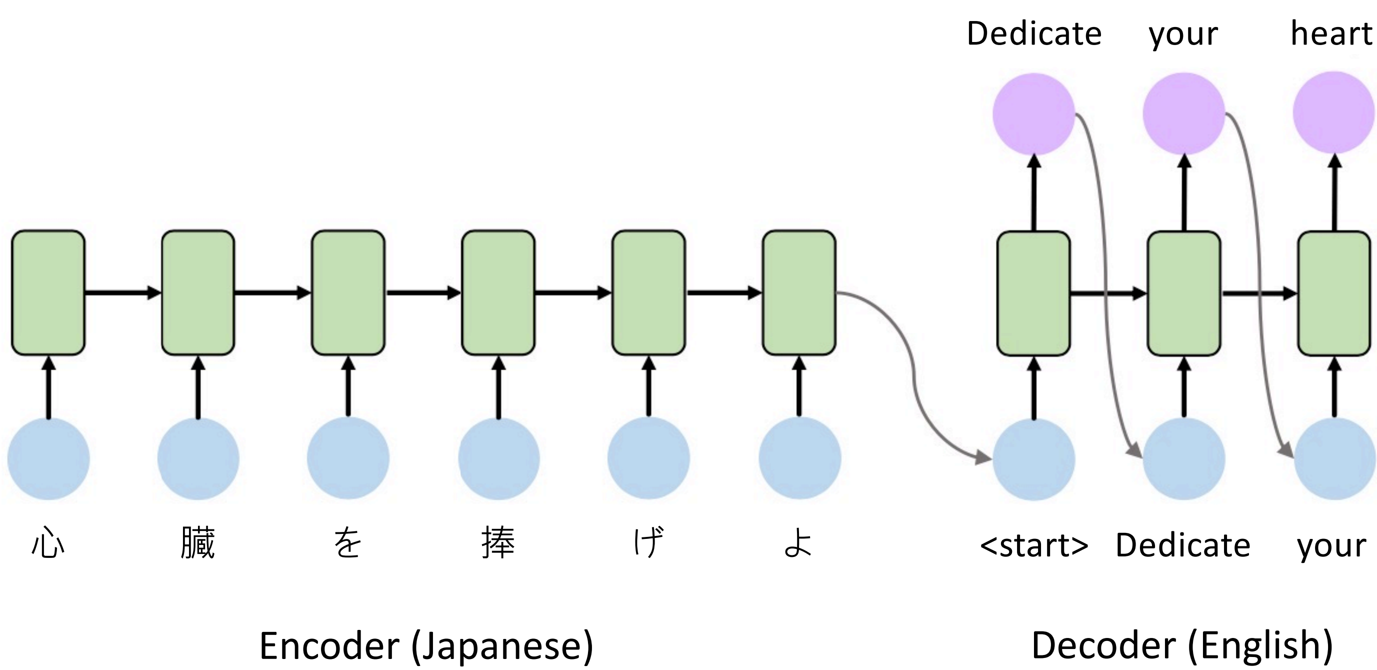


RNN Applications: Machine Translation



Attack on Titan, Source: https://www.youtube.com/watch?v=BhipGqSZEB0

RNN Applications: Machine Translation

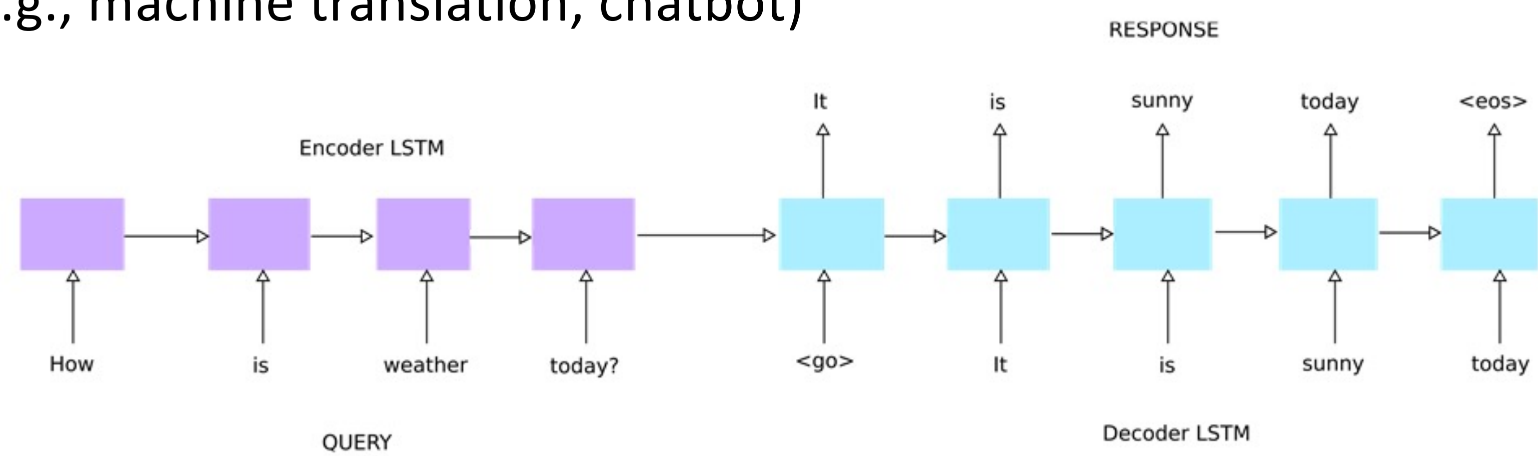


Sequence-to-sequence [1]
(Seq2Seq)

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

Sequence-to-Sequence Learning (Seq2Seq)

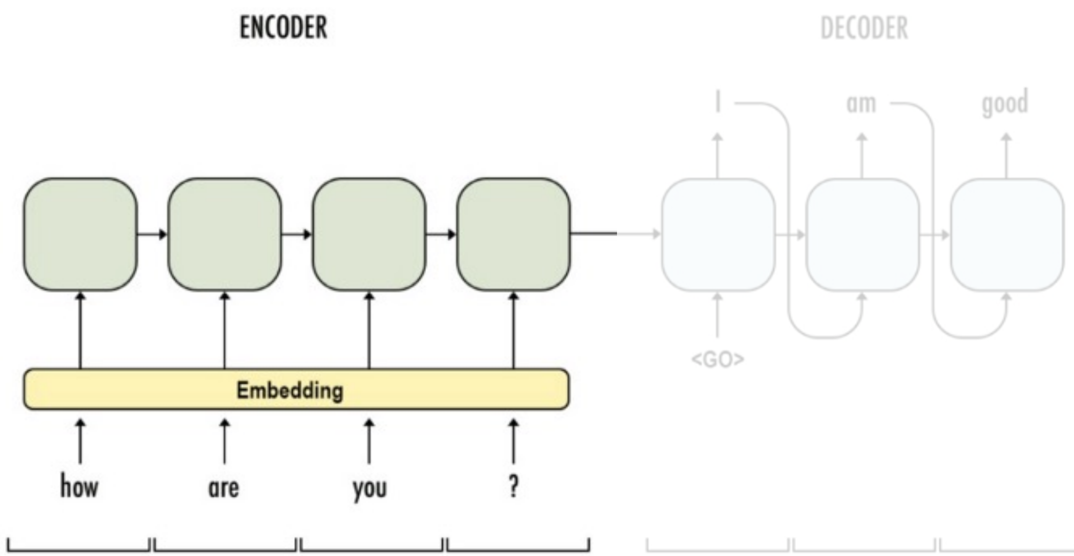
- Seq2Seq is to train models that convert **sequences from one domain** (e.g., sentences in Japanese) to **sequences in another domain** (e.g., the same sentences translated to English)
- Input sequences and output sequences can **have different lengths** (e.g., machine translation, chatbot)



Source: https://www.oreilly.com/library/view/deep-learning-essentials/9781785880360/b71e37fb-5fd9-4094-98c8-04130d5f0771.xhtml

Seq2Seq

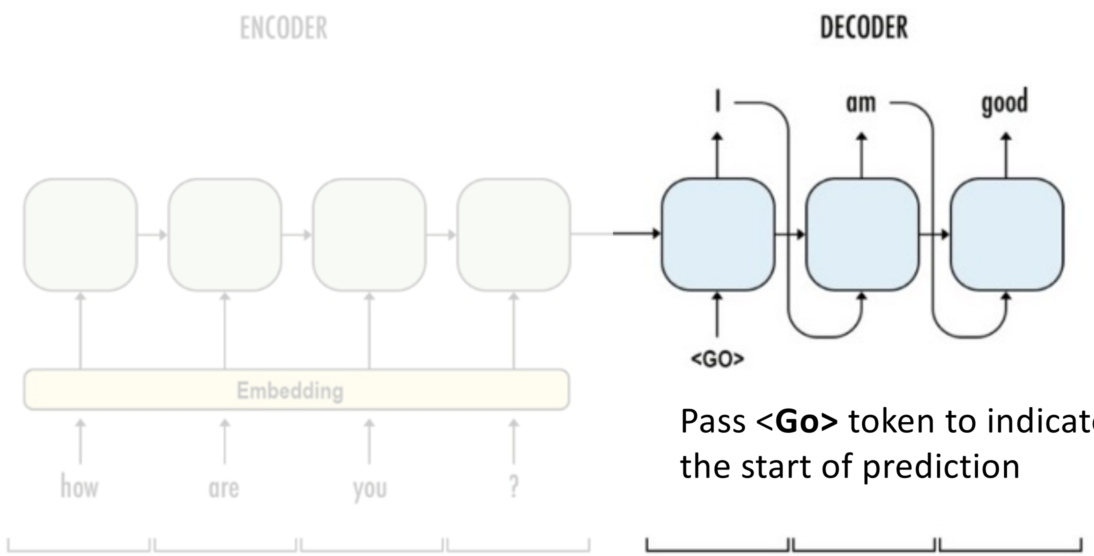
- The encoder transforms an input sequence of variable length into a fixed-shape context variable (e.g., hidden and cell states in LSTM)



Source: https://medium.com/analytics-vidhya/intuitive-understanding-of-seq2seq-model-attention-mechanism-in-deep-learning-1c1c24aace1e

Seq2Seq

- The decoder model is trained to predict the next word in the sequence given the previous word



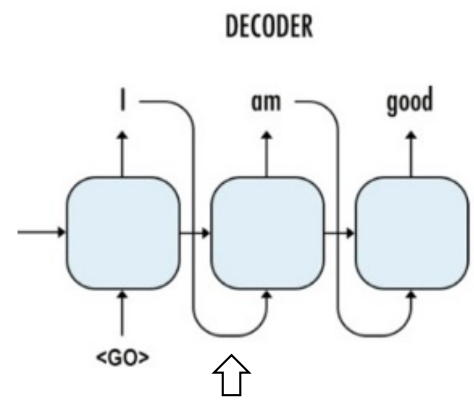
Pass <**Go**> token to indicate the start of prediction

Source: https://medium.com/analytics-vidhya/intuitive-understanding-of-seq2seq-model-attention-mechanism-in-deep-learning-1c1c24aace1e

Limitations of Seq2Seq I

- At the early stages of training, the predictions of the decoder are very bad
- The hidden states of the model will be updated by a sequence of wrong predictions, and **errors will accumulate**

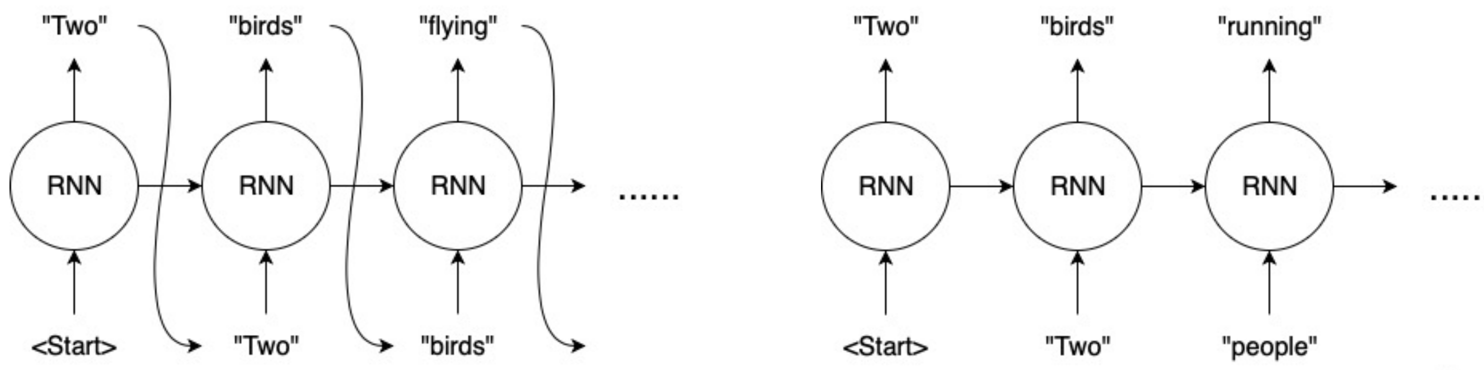
- **Solution: Teacher Forcing**



Feed with prediction from last step

Teacher Forcing

- Teacher forcing is a strategy for training recurrent neural networks that uses **ground truth as input**, instead of model output from a prior time step as an input



Assuming the ground truth sentence is “Two people running …”. Left figure: without Teacher Forcing, the model keeps feeding wrong word after making one mistake. Right figure: with Teacher Forcing, our model feeds “people” for the 3rd prediction.

Pros and Cons of Teacher Forcing

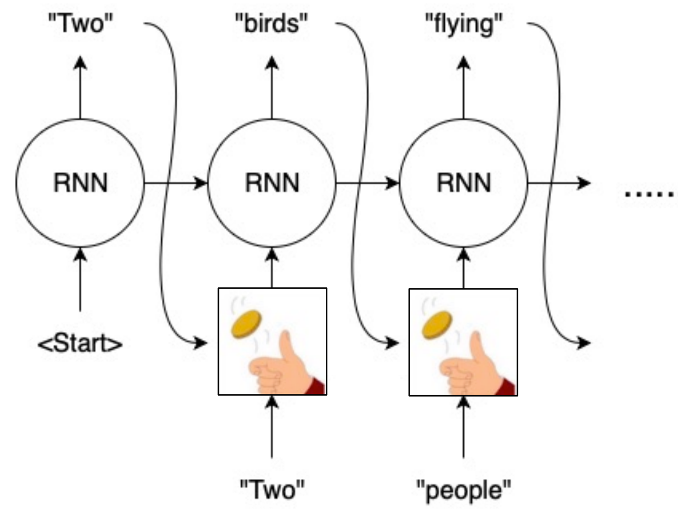
- Pros
  - Training with Teacher Forcing converges faster
  - Teacher Forcing can prevent error accumulation during training

- Cons
  - During inference, since there is no ground truth available, the RNN model will need to feed its own previous prediction for the next prediction
  - There is a discrepancy between training and inference, and might lead to poor model performance and instability

Curriculum Learning

- The curriculum learning is to randomly choose to use the **ground truth output** or **the generated output from the previous time step** as the input for the current time step

- The curriculum learning encourages the model to learn how to correct its own mistakes



Flip a coin to decide to use the true previous token or predicted token from the model

Scheduled Sampling

- The curriculum changes over time is called scheduled sampling [1]

- The scheduled sampling changes the training process **from a fully guided scheme** using the true previous token, **towards a less guided scheme** which mostly uses the generated token instead

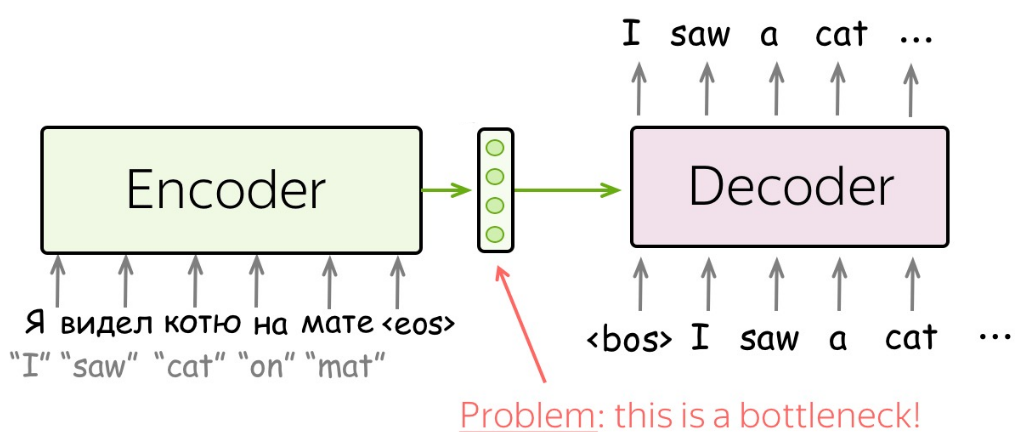
- For example, at the beginning of the training process, using Teacher Forcing. After several epochs, using the prediction as the input.

[1] Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, *28*.

Limitations of Seq2Seq II

- The encoder and decoder works fine for short sequence
- However, when the sequence is long, the encoder might be difficult to memorize the entire sequence into a fixed-sized vector and to compress all the contextual information in the sequence

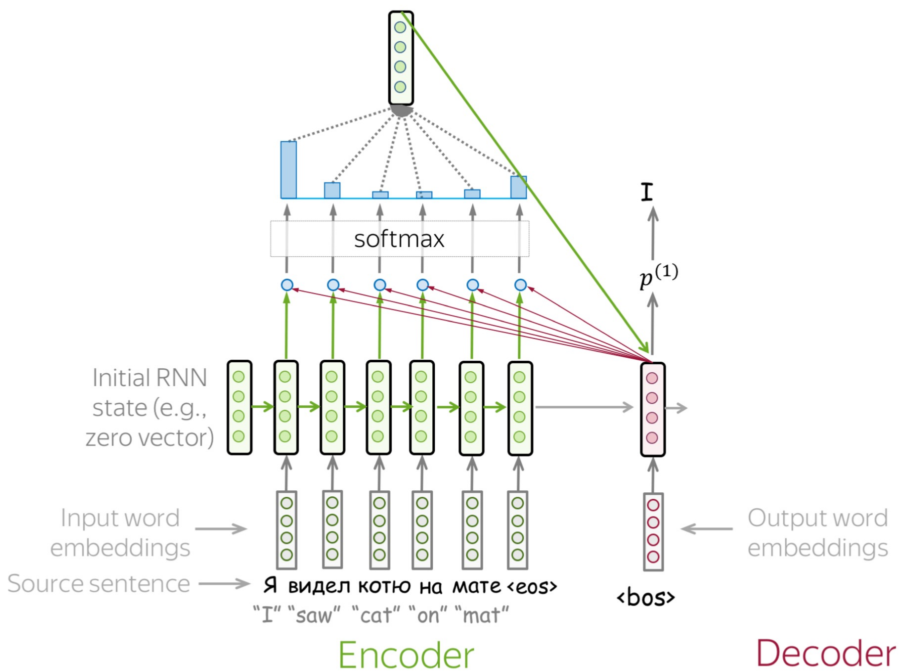
- **Solution: Attention**



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention in Seq2Seq

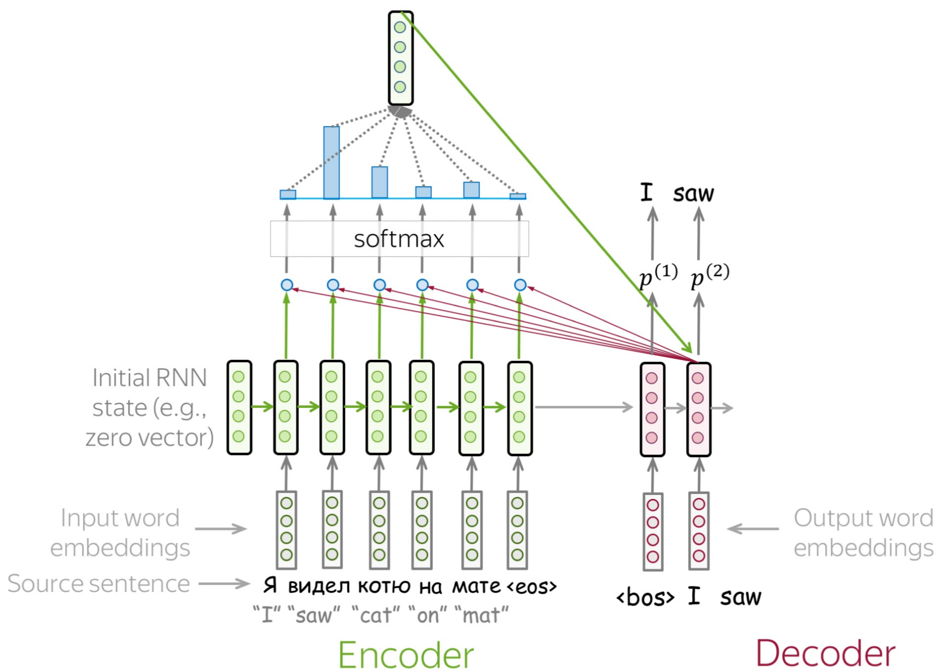
- Let the model "focus" on different parts of the output from encoder



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention in Seq2Seq

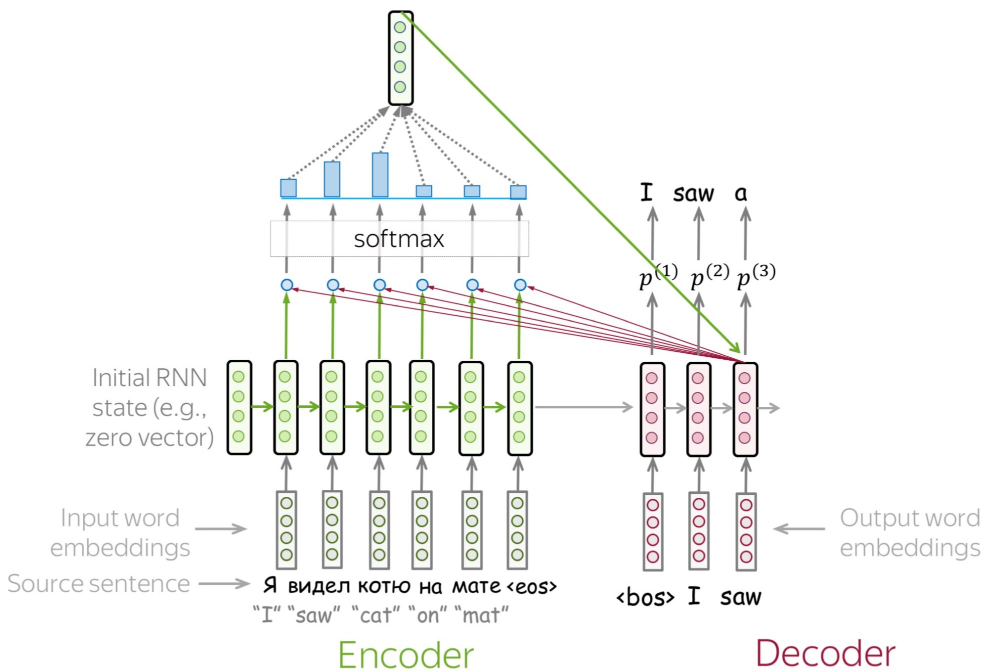
- Let the model "focus" on different parts of the output from encoder



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention in Seq2Seq

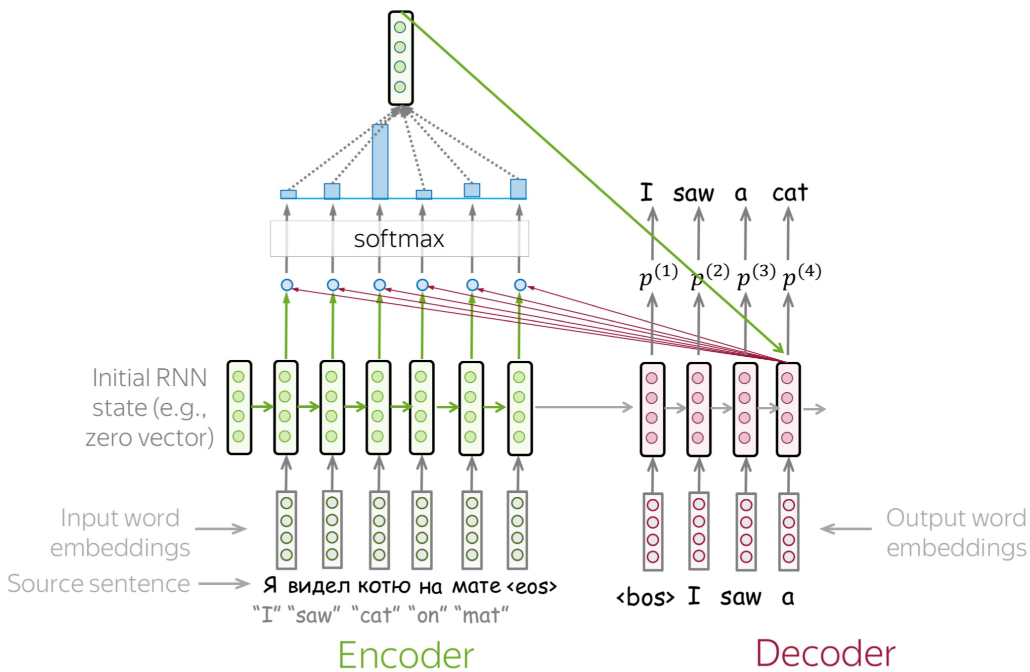
- Let the model "focus" on different parts of the output from encoder



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention in Seq2Seq

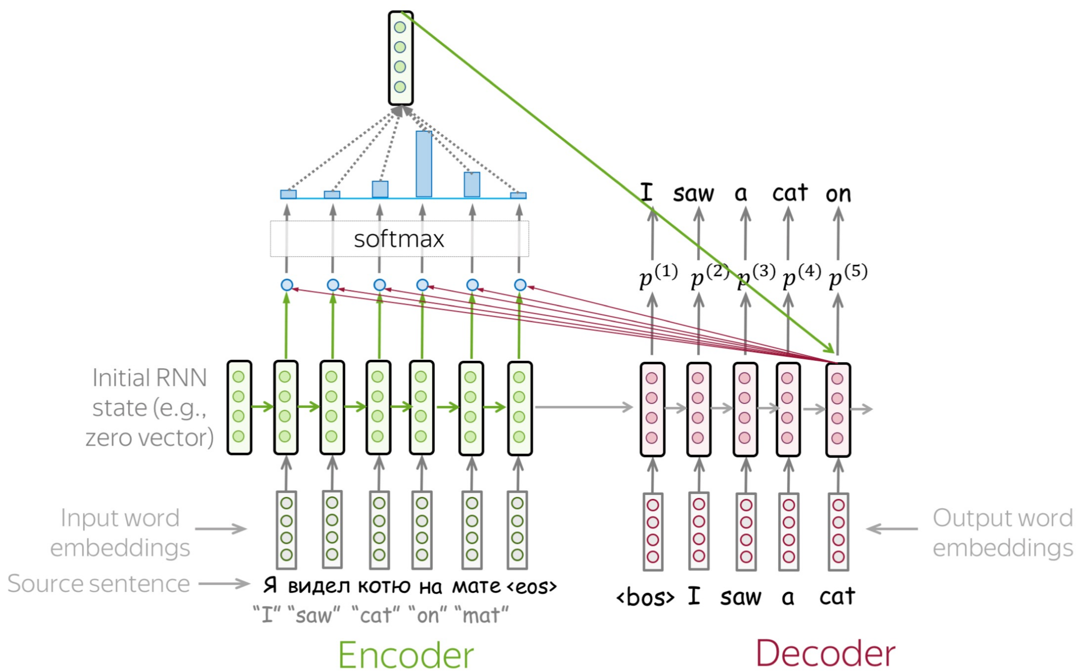
- Let the model "focus" on different parts of the output from encoder



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Attention in Seq2Seq

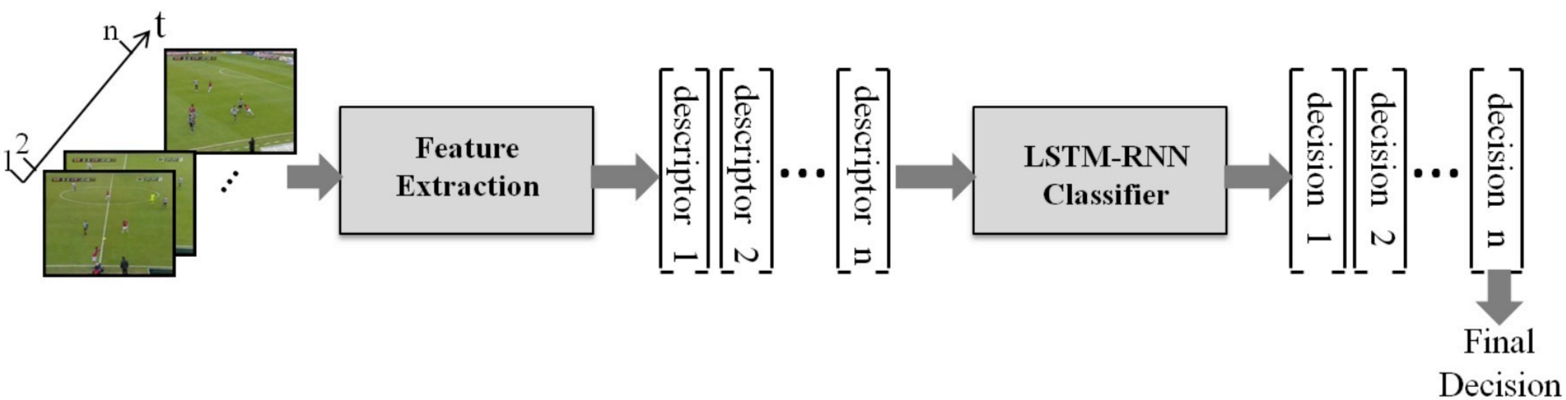
- Let the model "focus" on different parts of the output from encoder



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

More than Language Models
RNN in Sports

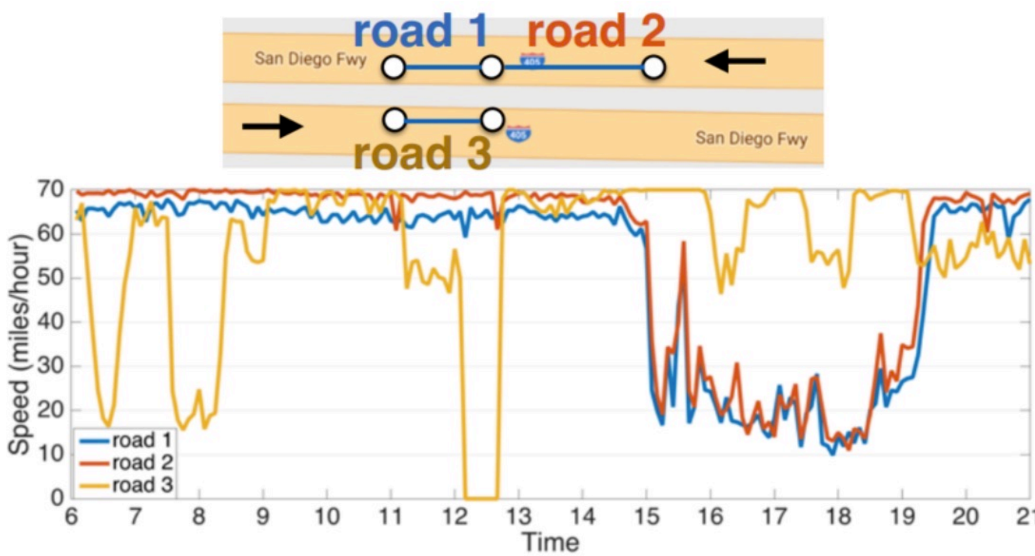
- Sport is a sequence of event (e.g., sequence of images, voices)
  - Applying RNN to basketball trajectories in the form of sequence modeling to predict whether a three-point shot is successful [1]
  - Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks [2]



[1] Shah, Rajiv, and Rob Romijnders. "Applying Deep Learning to Basketball Trajectories." arXiv preprint arXiv:1608.03793 (2016).
[2] Baccouche, Moez, et al. "Action classification in soccer videos with long short-term memory recurrent neural networks." International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2010. (The Image is taken from the paper)

More than Language Model
Traffic Forecasting

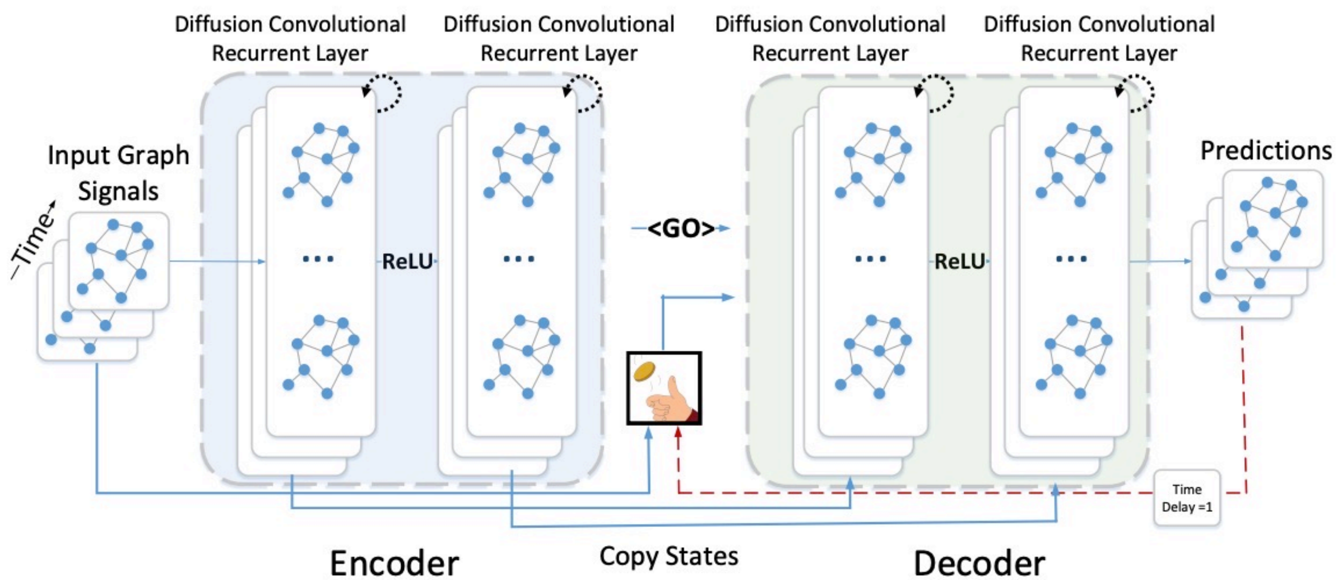
- Traffic forecasting is to predict the future traffic speeds of a sensor network given historic traffic speeds and the underlying road networks



[1] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926. Images are taken from the paper.

Traffic Forecasting

- The paper [1] constructs the sensor network as a graph and embeds graph in the RNN model to capture spatiotemporal evolution



[1] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926. Images are taken from the paper.

Summary

- GRU is another RNN variant
- Stacked RNN and bidirectional RNN are useful RNN architectures
- Teaching forcing and curriculum learning strategies accelerate the training process and improve the overall performance
- RNNs have been applied in various domains and applications combined with other techniques
  - Convolution Neural Networks -> **ConvLSTM**
  - Graph convolution
  - Attention

Acknowledgements

- Deep learning slides adapted from https://m2dsupsdlclass.github.io/lectures-labs/ by Olivier Grisel and Charles Ollion (CC-By 4.0 license)
- Gil, Yolanda (Ed.) Introduction to Computational Thinking and Data Science. Available from http://www.datascience4all.org
- https://lilianweng.github.io/posts/2018-08-12-vae/
- https://towardsdatascience.com/rnn-cells-analyzing-gru-equations-vs-lstm-and-when-to-choose-rnn-over-transformers-a28e46beef31
- https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html