# Outline

# Outline

# Tasks

- Question Answering
- Summarization
- Query Suggestion
- Reading Comprehension / Wiki Reading
- Dialogue Systems
  - Goal-Oriented
  - Chit-Chat

# Example Scenario for machine reading task

Sandra went to the kitchen. Fred went to the kitchen. Sandra picked up the milk. Sandra traveled to the office. Sandra left the milk. Sandra went to the bathroom.

- Where is the milk now? A: office
- Where is Sandra? A: bathroom
- Where was Sandra before the office? A: kitchen

# Example Scenario for machine reading task

Sandra went to the kitchen. Fred went to the kitchen. Sandra picked up the milk. Sandra traveled to the office. Sandra left the milk. Sandra went to the bathroom.

- Where is the milk now? A: office
- Where is Sandra? A: bathroom
- Where was Sandra before the office? A: kitchen

I'll be going to Los Angeles shortly. I want to book a flight. I am leaving from Amsterdam. I want the return flight to be early morning. I don't have any extra luggage. I wouldn't mind extra leg room.

- What does the user want? A: Book a flight
- Where is the user flying from? A: Amsterdam
- Where is the use going to? A: Los Angeles

# What is Required?

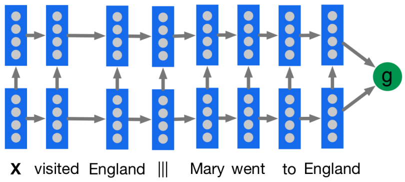
- The model needs to remember the context
- It needs to know what to look for in the context
- Given an input, the model needs to know where to look in the context
- It needs to know how to reason using this context
- It needs to handle changes in the context

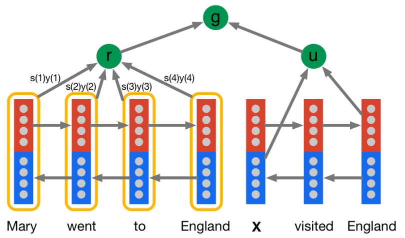
A Possible Solution:

- Hidden states of RNNs have memory: Run an RNN on the and get its representation to map question to answers/response.

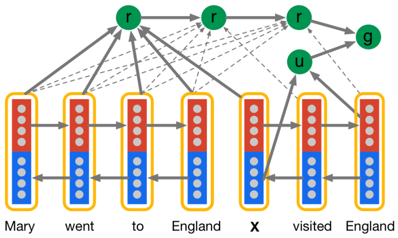
This will not scale as RNN states don't have ability to capture long term dependency: vanishing gradients, limited state size.

# Teaching Machine to Read and Comprehend




(a) Attentive Reader.

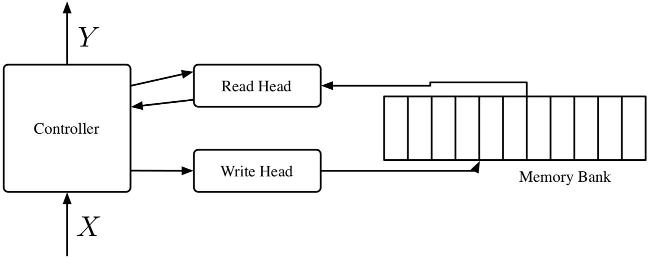



(b) Impatient Reader.

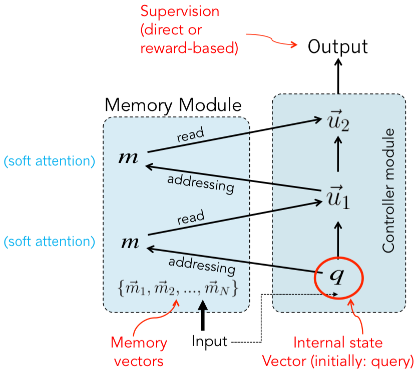
[Hermann et al., 2015]

# Neural Networks with Memory

- Memory Networks
  - End2End MemNNs
  - Key-Value MemNNs
- Neural Turing Machines
- Stack/List/Queue Augmented RNNs

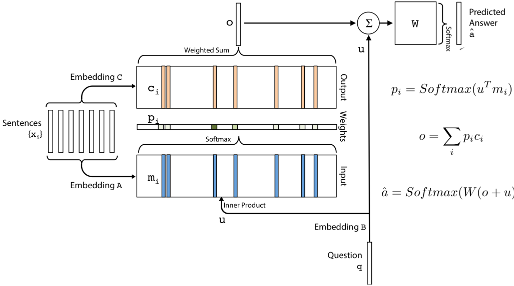


# End2End Memory Networks [Sukhbaatar et al., 2015]




- Learns which parts of the memory are relevant.
- This is achieved by reading using attention.
- Performs multiple lookups to refine its guess about memory relevance.
- Only needs supervision at the final output.

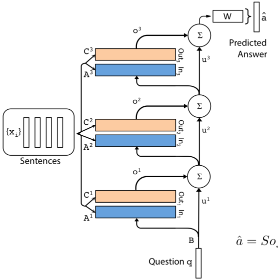
# End2End Memory Networks (Multiple Hops)




$$p_i = Softmax(u^T m_i)$$

$$o = \sum_i p_i c_i$$

$$\hat{a} = Softmax(W(o + u))$$




$${}^k p_i = Softmax({}^k u^T \cdot {}^k m_i)$$

$${}^k o = \sum_i {}^k p_i {}^k c_i$$

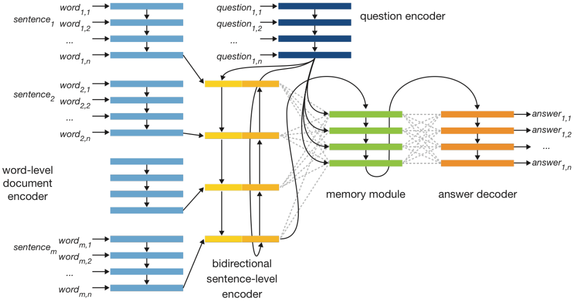
$${}^{k+1} u = {}^k u + {}^k o$$

$$\hat{a} = Softmax(W \cdot {}^{k+1} u) = Softmax(W({}^k o + {}^k u))$$

- Share the input and output embeddings or not
- What to store in memories individual words, word windows, full sentences
- How to represent the memories? Bag-of-words? RNN reading of words? Characters?

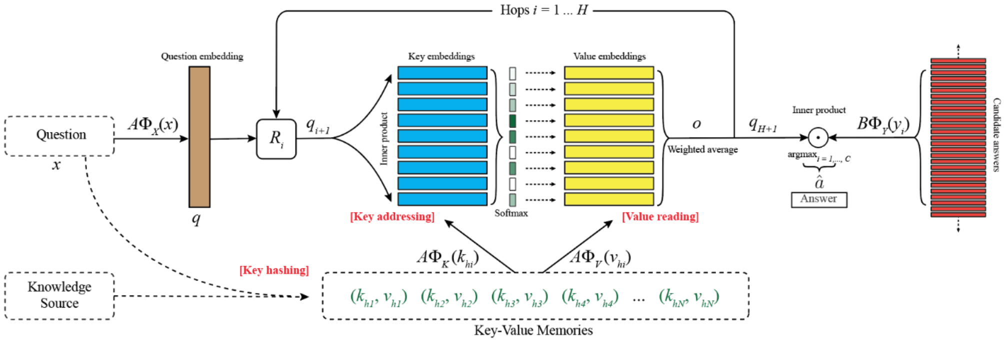
# Attentive Memory Networks [Kenter and de Rijke, 2017]

- Proposed Model: An end-to-end trainable memory networks with a hierarchical input encoder.




- Framing the task of conversational search as a general machine reading task.

# Key-Value Memory Networks




## Example:

for a KB triple [subject, relation, object], Key could be [subject,relation] and value could be [object] or vice versa.

[Miller et al., 2016]

# WikiReading [Hewlett et al., 2016, Kenter et al., 2018]

Task is based on Wikipedia data (datasets available in English, Turkish and Russian).

| | Categorization | | Extraction | |
|---|---|---|---|---|
| **Document** | Folkart Towers are twin skyscrapers in the Bayrakli district of the Turkish city of Izmir. Reaching a structural height of 200 m (656 ft) above ground level, they are the tallest . . . | Angeles blancos is a Mexican telenovela produced by Carlos Sotomayor for Televisa in 1990. Jacqueline Andere, Rogelio Guerra and Alfonso Iturralde star as the main . . . | Canada is a country in the northern part of North America. Its ten provinces and three territories extend from the **Atlantic** to the **Pacific** and northward into the **Arctic Ocean**, . . . | Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. The show originally aired on the AMC network for five seasons, from **January 20, 2008**, to . . . |
| **Property** | country | original language of work | located next to body of water | start time |
| **Answer** | Turkey | Spanish | Atlantic Ocean, Arctic Ocean, Pacific Ocean | 20 January 2008 |

- Categorical: relatively small number of possible answer (e.g.: instance of, gender, country).
- Relational: rare or totally unique answers (e.g.: date of birth, parent,capital).

# WikiReading

- **Answer Classification**: Encoding document and question, using softmax classifier to assign probability to each of to-50k answers (limited answer vocab) .
  - Sparse BoW Baseline, Averaged Embeddings, Paragraph Vector, LSTM Reader, Attentive Reader, Memory Network.
  - Generally models with RNN and attention work better, especially at relational properties.
- **Answer Extraction** (labeling/pointing) For each word in the document, compute the probability that it is part of the answer.
  - Regardless of the vocabulary so the answer requires being mentioned in the document.
  - RNN Labeler: shows a complementary set of strengths, performing better on relational properties than categorical ones
- **Sequence to Sequence** Encoding query and document and decoding the answer as sequences of words or characters.
  - Basic seq2seq, Placeholder seq2seq, Basic Character seq2seq,
  - Unifies the classification and extraction in one model: Greater degree of balance between relational and categorical properties.

# Outline

# Dialogue systems

## Dialogues/conversational agents/chat bots

Open-ended dialogues

- ELIZA
- Twitterbots
- Alexa/Google home/Siri/Cortana

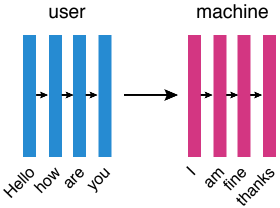
Goal-oriented dialogues

- Restaurant finding
- Hotel reservations
- Set an alarm clock
- Order a pizza
- Play music
- Alexa/Google home/Siri/Cortana

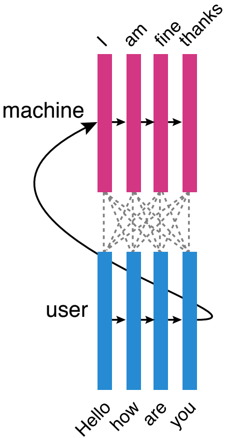
Is this IR?

# Dialogue systems

## Chit-chat bots




Straightforward seq-to-seq [Vinyals and Le, 2015].
([Sordoni et al., 2015b] is a precursor, but no RNN-to-RNN, and no LSTM).




Same idea, but with attention [Shang et al., 2015]

# Dialogue systems

## Limitations

- 'Wrong' optimization criterion
- Generic responses
- No way to incorporate world knowledge
- No model of conversation
  - Inconsistency
  - No memory of what was said earlier on

**Human**: *what is your job?*
**Machine**: *i'm a lawyer.*
**Human**: *what do you do?*
**Machine**: *i'm a doctor.*
Example from [Vinyals and Le, 2015]

## Evaluation

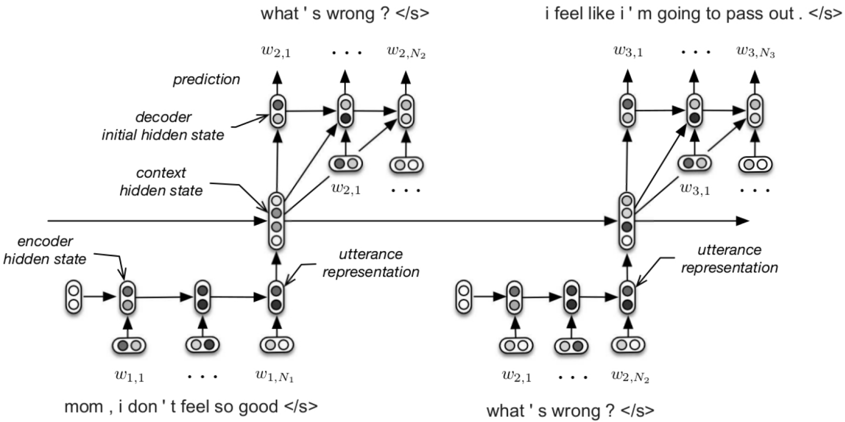
- Perplexity?
- BLUE/METEOR?
- Nice overview of *How NOT To Evaluate Your Dialogue System* [Liu et al., 2016].
- Open problem....

# Dialogue systems

## 3 solutions

- More consistency in dialogue with hierarchical network
- Less generic responses with different optimization function
- More natural responses with GANs

# Dialogue systems




Hierarchical seq-to-seq [Serban et al., 2016]. Main evaluation metric: perplexity.

# Dialogue systems

## Avoid generic responses

Usually: optimize log likelihood of predicted utterance, given previous context:

$$C_{LL} = \arg\max_{u_t} log\, p(u_t|\textsf{context}) = \arg\max_{u_t} log\, p(u_t|u_0 \ldots u_{t-1})$$

To avoid repetitive/boring answer (*I don't know*), use maximum mutual information between previous context and predicted utterance [Li et al., 2015].

$$\begin{aligned} C_{MMI} &= \arg\max_{u_t} log \frac{p(u_t, \textsf{context})}{p(u_t)p(\textsf{context})} \\ &= \textit{[derivation, next page \ldots]} \\ &= \arg\max_{u_t} (1-\lambda) log\, p(u_t|\textsf{context}) + \lambda log\, p(\textsf{context}|u_t) \end{aligned}$$

# Dialogue systems

Bayes rule

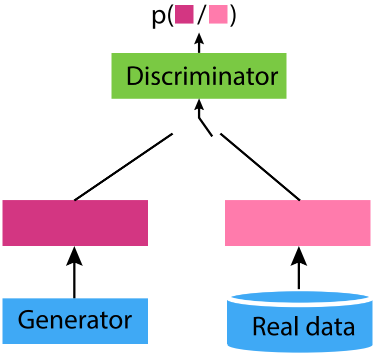
$$\begin{aligned}
log\, p(u_t|\textit{context}) &= log\, \frac{p(\textit{context}|u_t)p(u_t)}{p(\textit{context})} \\
log\, p(u_t|\textit{context}) &= log\, p(\textit{context}|u_t) + log\, p(u_t) - log\, p(\textit{context}) \\
log\, p(u_t) &= log\, p(u_t|\textit{context}) - log\, p(\textit{context}|u_t) + log\, p(\textit{context})
\end{aligned}$$

$$\begin{aligned}
C_{MMI} &= \arg\max_{u_t} log\, \frac{p(u_t, \textit{context})}{p(u_t)p(\textit{context})} = \arg\max_{u_t} log\, \frac{p(u_t|\textit{context})p(\textit{context})}{p(u_t)p(\textit{context})} \\
&= \arg\max_{u_t} log\, \frac{p(u_t|\textit{context})}{p(u_t)} \\
&= \arg\max_{u_t} log\, p(u_t|\textit{context}) - log\, p(u_t) \leftarrow \text{Weird, minus language model score.} \\
&= \arg\max_{u_t} log\, p(u_t|\textit{context}) - \lambda log\, p(u_t) \leftarrow \text{Introduce } \lambda\text{. Crucial step! Without this it wouldn't work.} \\
&= \arg\max_{u_t} log\, p(u_t|\textit{context}) - \lambda(log\, p(u_t|\textit{context}) - log\, p(\textit{context}|u_t) + log\, p(\textit{context})) \\
&= \arg\max_{u_t} (1-\lambda) log\, p(u_t|\textit{context}) + \lambda log\, p(\textit{context}|u_t)
\end{aligned}$$

*(More is needed to get it to work. See [Li et al., 2015] for more details.)*

# Generative adversarial network for dialogues

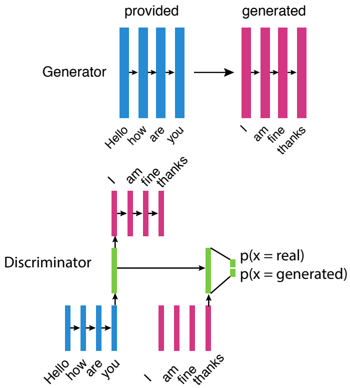
- Discriminator network
  - Classifier: real or generated utterance
- Generator network
  - Generate a realistic utterance




Original GAN paper [Goodfellow et al., 2014].
Conditional GANs, e.g. [Isola et al., 2016].

# Generative adversarial network for dialogues

- Discriminator network
  - Classifier: real or generated utterance
- Generator network
  - Generate a realistic utterance




See [Li et al., 2017] for more details.

Code available at `https://github.com/jiweil/Neural-Dialogue-Generation`

# Dialogue systems

## Open-ended dialogue systems

- Very cool, current problem
- Very hard
- Many problems
  - Training data
  - Evaluation
  - Consistency
  - Persona
  - ...

# Outline

# Goal-oriented

## Idea

- Closed domain
  - Restaurant reservations
  - Finding movies
- Have a dialogue system find out what the user wants

## Challenges

- Training data
- Keeping track of dialogue history
- Handling of out-of-domain words or requests
- Going beyond task-specific slot filling
- Intermingling live API calls, chit chat, information requests, etc.
- Evaluation
  - Solve the task
  - Naturalness
  - Tone of voice
  - Speed
  - Error recovery

# Goal-oriented as seq2seq

## Memory network [Bordes and Weston, 2017]

- Simulated dataset
- Finite set of things the bot can say
  - Because of the way the dataset is constructed
- Memory networks
- Training: next utterance prediction
- Evaluation
  - response-level
  - dialogue-level

Restaurant Knowledge Base, i.e., a table.
Queried by API calls.
Each row = restaurant:

- cuisine (10 choices, e.g., French, Thai)
- location (10 choices, e.g., London, Tokyo)
- price range (cheap, moderate or expensive)
- rating (from 1 to 8)

For words of relevant entity types

- add a trainable entity vector

# Goal-oriented as reinforcement learning

A typical reinforcement learning system:

- States $S$
- Actions $A$
- State transition function: $T : S, A \rightarrow S$
- Reward function: $R : S, A, S \rightarrow \mathbb{R}$
- Policy: $\pi : S \rightarrow A$

A RL system needs an environment to interact with (e.g., real users).

Typically [Shah et al., 2016]:

- States: agents interpretation of the environment: distribution over user intents, dialogue acts and slots and their values
  - intent(buy_ticket)
  - inform(destination=Atlanta)
  - ...
- Actions: possible communications, and are usually designed as a combination of dialogue act tags, slots and possibly slot values
  - request(departure_date)
  - ...

# Goal-oriented as reinforcement learning

Restaurant finding [Wen et al., 2017]:

- Neural belief tracking: distribution over a possible values of a set of slots
- Delexicalisation: swap slot-values for generic token (e.g. *Chinese*, *Indian*, *Italian* → *FOOD_TYPE*)

Movie finding [Dhingra et al., 2017]:

- Simulated user
- Soft attention over database
- Neural belief tracking:
  - Multinomial distribution for every column over possible column values
  - RNN, input is dialogue so far, output softmax over possible column values

Reward based on finding the right KB entry.

# Goal-oriented

## Goal-oriented models

- Currently works primarily in very small domains
- How about multiple speakers?
- Not clear what kind of architecture is best
- Reinforcement learning might be the way to go (?)
- Open research area...

# Outline

# Alternatives to RNNs

RNNs are:

- Well-studied
- Robust and tried and trusted method for sequence tasks

However, RNNs have several drawbacks:

- Take time to train
- Expensive to unroll for many steps
- Not too good at catching long-term dependencies

Can we do better?

- WaveNet
- ByteNet
- Transformer

# Alternatives to RNNs: WaveNet

WaveNet is originally introduced for a text-to-speech task (i.e. generating realistic audio waves).
We try to model:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}).$$

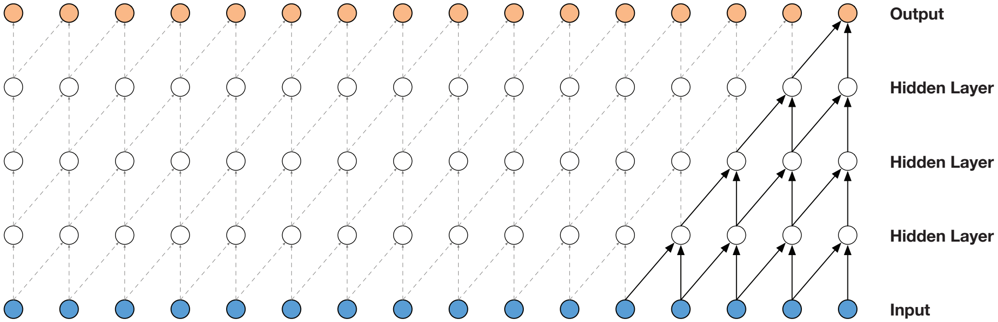
- Stack of convolutional layers. No pooling layers.
- Output of the model has the same time dimensionality as the input.
- Output is a categorical distribution over the next value $x_t$ with a softmax layer and it is optimized to maximize the log-likelihood of the data w.r.t. the parameters.

Based on the idea of dilated causal convolutions.

[van den Oord et al., 2016]

# Alternatives to RNNs: WaveNet

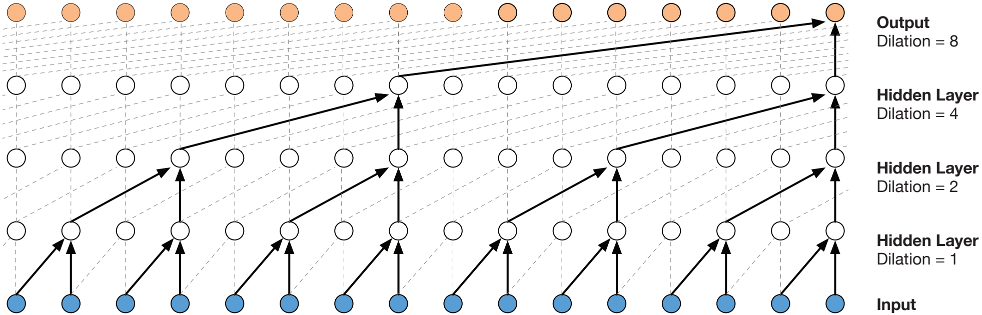
Causal convolutions




[van den Oord et al., 2016]

# Alternatives to RNNs: WaveNet

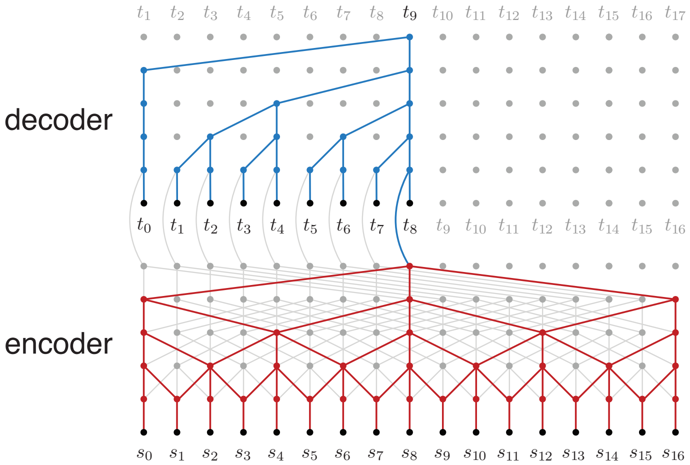
## Dilated causal convolutions




*"At training time, the conditional predictions for all timesteps can be made in parallel because all timesteps of ground truth x are known. When generating with the model, the predictions are sequential: after each sample is predicted, it is fed back into the network to predict the next sample."*

[van den Oord et al., 2016]

# Alternatives to RNNs: ByteNet



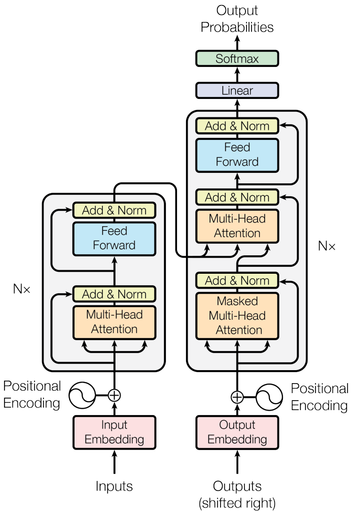

[Kalchbrenner et al., 2016]

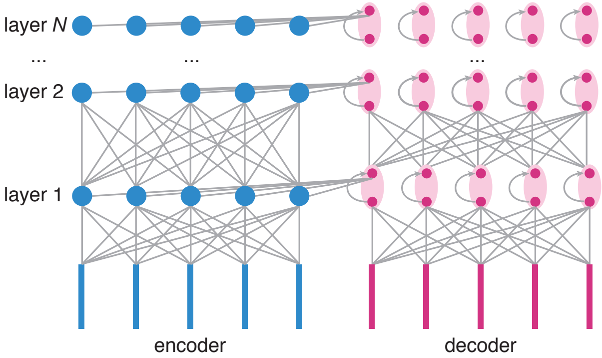
# Alternatives to RNNs: Transformer

- Positional encoding added to the input embeddings
- Key-value attention
- Multi-head self-attention
- The encoder attends over its own states
- The decoder alters between
  - attending over its own inputs/states
  - attending over encoder states at the same level

[Vaswani et al., 2017]

# Alternatives to RNNs: Transformer



# Outline

# Resources: datasets

## Open-ended dialogue

- Opensubtitles [Tiedemann, 2009]
- Twitter: http://research.microsoft.com/convo/
- Weibo:
http://www.noahlab.com.hk/topics/ShortTextConversation
- Ubuntu Dialogue Corpus [Lowe et al., 2015]
- Switchboard
https://web.stanford.edu/~jurafsky/ws97/
- Coarse Discourse (Google Research)
https://research.googleblog.com/2017/05/coarse-discourse-dataset-for.html

## Goal-oriented dialogues

- MISC: A data set of information-seeking conversations [Thomas et al., 2017]
- Maluuba Frames
http://datasets.maluuba.com/Frames
- Loqui Human-Human Dialogue Corpus
https://academiccommons.columbia.edu/catalog/ac:176612
- bAbi (Facebook Research)
https://research.fb.com/downloads/babi/

## Machine reading

- bAbi QA (Facebook Research)
https://research.fb.com/downloads/babi/
- QA Corpus [Hermann et al., 2015]
https://github.com/deepmind/rc-data/
- WikiReading (Google Research)
https://github.com/google-research-datasets/wiki-reading

# Resources: source code

- End-to-end memory network

  https://github.com/facebook/MemNN

- Attentive Memory Networks

  https://bitbucket.org/TomKenter/attentive-memory-networks-code

- Hierarchical NN [Serban et al., 2016]

  https://github.com/julianser/hed-dlg, https://github.com/julianser/rnn-lm

- GAN for dialogues

  https://github.com/jiweil/Neural-Dialogue-Generation

- RL for dialogue agents [Dhingra et al., 2017]

  https://github.com/MiuLab/KB-InfoBot

- Transformer network

  https://github.com/tensorflow/tensor2tensor