Breaking news!

• Practical 2 is due April 6, 23:59

• Midterm exam in class next week (April 12)
  – Check the midterm guide for details

• Practical 3 will be out!
  – Language modeling with RNNs
  – Due Sunday, April 27, 23:59

PANDARUS:
Alas, I think he shall be come approached and the day
When little sain would be attain’d into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.
Previously on CMP784

- Sequence modeling
- Recurrent Neural Networks (RNNs)
- The Vanilla RNN unit
- How to train RNNs
- The Long Short-Term Memory (LSTM) unit and its variants
- Gated Recurrent Unit (GRU)
Lecture overview

• Attention Mechanism for Deep Learning
• Attention for Image Captioning
• Memory Networks
• End-to-end Memory Networks
• Dynamic Memory Networks

Disclaimer: Much of the material and slides for this lecture were borrowed from
— Mateusz Malinowski’s lecture on Attention-based Networks
— Graham Neubig’s CMU CS11-747 Neural Networks for NLP class
— Chris Dyer’s Oxford Deep NLP class
— Yoshua Bengio’s talk on From Attention to Memory and towards Longer-Term Dependencies
— Sumit Chopra’s lecture on Reasoning, Attention and Memory
— Jason Weston’s tutorial on Memory Networks for Language Understanding
— Richard Socher’s talk on Dynamic Memory Networks
Deep Learning for Vision

What if we treat an existing deep model as a black box in pedestrian detection?

ConvNet−U−MS


Figure credit: Xiaogang Wang
Deep Learning for Speech

“He can for example present significant university wide issues to the senate.”

Spectrogram: window in time $\rightarrow$ vector of frequencies; slide; repeat

Figure credit: NVidia
“The movie was not bad at all. I had fun.”
Deep Models

Input Representation

$G_{W_2}$
Classifier/Regressor (decoder)

Typically a Linear Projection with some non-linearity (log-soft-max)

$F_{W_1}$
Feature Extractor (encoder)

Fully Connected Network
Convolution Network
Recurrent Network

"The movie was not bad at all. I had fun."

can be seen as a prior on the type of transformation you want
Deep Models

Learnable parametric function

Inputs: generally considered I.I.D.

Outputs: classification or regression

“The movie was not bad at all. I had fun.”
Encoder-Decoder Framework

- Intermediate representation of meaning
  = ‘universal representation’
- Encoder: from word sequence to sentence representation
- Decoder: from representation to word sequence distribution
Sentence Representations

• But what if we could use multiple vectors, based on the length of the sentence.

  this is an example →

  this is an example →
Attention
Basic Idea

• Encode each word in the sentence into a vector

• When decoding, perform a linear combination of these vectors, weighted by “attention weights” (where to look)

• Use this combination in picking the next item
Calculating Attention

• Use **query** vector (decoder state) and **key** vectors (all encoder states)

• For each query-key pair, calculate weight

• Normalize to add to one using softmax
Calculating Attention

• Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum

\[ \alpha_1 = 0.76 \quad \alpha_2 = 0.08 \quad \alpha_3 = 0.13 \quad \alpha_4 = 0.03 \]

• Use this in any part of the model you like
A Graphical Example

End-to-End Machine Translation with Recurrent Nets and Attention Mechanism

Attention Score Functions

• \( q \) is the query and \( k \) is the key

• Multi-layer Perceptron (Bahdanau et al. 2015)
  \[
  a(q, k) = w_2^\top \tanh(W_1[q; k])
  \]
  - Flexible, often very good with large data

• Bilinear (Luong et al. 2015)
  \[
  a(q, k) = q^\top W k
  \]

• Dot Product (Luong et al. 2015)
  \[
  a(q, k) = q^\top k
  \]
  - No parameters! But requires sizes to be the same.

• Scaled Dot Product (Vaswani et al. 2017)
  - Problem: scale of dot product increases as dimensions get larger
  - Fix: scale by size of the vector
  \[
  a(q, k) = \frac{q^\top k}{\sqrt{|k|}}
  \]
Case Study: Show, Attend and Tell

Paying Attention to Selected Parts of the Image While Uttering Words
Vinyals et al. (2014) Show and Tell: A Neural Image Caption Generator
Regions in ConvNets

• Each point in a “higher” level of a convnet defines spatially localized feature vectors/matrices.
• Xu et al. calls these “annotation vectors”, $a_i$, $i \in \{1, \ldots, L\}$
Regions in ConvNets

\[ F = \begin{bmatrix} a_1 \end{bmatrix} \]
Regions in ConvNets

\[ F = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \]
Regions in ConvNets

\[ F = \begin{bmatrix} a_1 & a_2 & a_3 & \ldots \end{bmatrix} \]
Extension of LSTM via the context vector

- Extract L D-dimensional annotations
  - Lower convolutional layer to have the correspondence between the feature vectors and portions of the 2-D image

\[
\begin{align*}
\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} &= \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} Ey_{t-1} \\ h_{t-1} \\ z_t \end{pmatrix} \\
\text{c}_t &= f_t \odot \text{c}_{t-1} + i_t \odot g_t \\
\text{h}_t &= o_t \odot \tanh(\text{c}_t). \\
e_{ti} &= f_{\text{att}}(a_i, \text{h}_{t-1}) \\
\alpha_{ti} &= \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.
\end{align*}
\]

\[\hat{z}_t = \phi(\{a_i\}, \{\alpha_i\}) \quad \phi \text{ is the ‘attention’ (‘focus’) function – ‘soft’ / ‘hard’}\]

\[p(y_t | a, y_{1:t-1}) \propto \exp(L_o(Ey_{t-1} + L_h h_t + L_z \hat{z}_t))\]
Hard attention

We have two sequences
'1' that runs over localizations
't' that runs over words

Stochastic decisions are discrete here, so derivatives are zero

\[ e_{ti} = f_{\text{att}}(a_i, h_{t-1}) \]
\[ \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \]
\[ \hat{z}_t = \phi \left( \{a_i\}, \{\alpha_i\} \right) \]

Loss is a variational lower bound on the marginal log-likelihood
\[ L_s = \sum_s p(s \mid a) \log p(y \mid s, a) \]
\[ \leq \log \sum_s p(s \mid a)p(y \mid s, a) \]
\[ = \log p(y \mid a) \]

Due to Jensen’s inequality \( E[\log(X)] \leq \log(E[X]) \)

\[ \hat{s}_t \sim \text{Multinoulli}_L(\{\alpha_i\}) \]

\[ \frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{\partial \log p(y \mid \hat{s}^n, a)}{\partial W} + \log p(y \mid \hat{s}^n, a) \frac{\partial \log p(\hat{s}^n \mid a)}{\partial W} \right] \]
\[ \lambda_v (\log p(y \mid \hat{s}^n, a) - b) \frac{\partial \log p(\hat{s}^n \mid a)}{\partial W} + \lambda_e \frac{\partial H[\hat{s}^n]}{\partial W} \]

To reduce the estimator variance, entropy term \( H[s] \) and bias are added [1,2]

Hard attention

We have two sequences ‘l’ that runs over localizations ‘t’ that runs over words

Stochastic decisions are discrete here, so derivatives are zero

Loss is a variational lower bound on the marginal log-likelihood

\[ L_s = \sum_s p(s \mid a) \log p(y_{1:t} \mid s, a) \]

\[ \leq \log \sum_s p(s \mid a) p(y_{1:t} \mid s, a) \]

\[ = \log p(y_{1:t} \mid a) \]

Due to Jensen’s inequality

\[ \tilde{s}_t \sim \text{Multinoulli}_L(\{\alpha_i\}) \]

\[ \frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{\partial \log p(y_{1:t} \mid \tilde{s}^n, a)}{\partial W} + \log p(y_{1:t} \mid \tilde{s}^n, a) \frac{\partial \log p(\tilde{s}^n \mid a)}{\partial W} \right] \]

\[ e_{ti} = f_{\text{att}}(a_i, h_{t-1}) \]

\[ \exp(e_{ti}) \]

\[ \hat{z}_t = \phi(\{a_i\}, \{\alpha_i\}) \]

\[ \alpha_{ti} = \frac{1}{\sum_{k=1}^{L} \exp(e_{tk})} \]

- Instead of a soft interpolation, make a zero-one decision about where to attend

- Harder to train, requires methods such as reinforcement learning

To reduce the estimator variance, entropy term \( H(s) \) and bias are added [1,2]

Soft attention

\[ \hat{z}_t = \sum_i s_{t,i} a_i \]

Instead of making hard decisions, we take the expected context vector

\[ \mathbb{E}_{p(s_t|a)}[\hat{z}_t] = \sum_{i=1}^{L} \alpha_{t,i} a_i \]

The whole model is smooth and differentiable under the deterministic attention; learning via a standard backprop

\[ \phi(\{a_i\}, \{\alpha_i\}) = \sum_i \alpha_i a_i \]

---

Theoretical arguments

- \[ \mathbb{E}_{p(s_t|a)}[h_t] \] equals to computing \( h_t \) using a single forward prop with the expected context vector \( \mathbb{E}_{p(s_t|a)}[\hat{z}_t] \)
- Normalized Weighted Geometric Mean approximation [1] \( NWGM[p(y_t = k \mid a)] \approx \mathbb{E}[p(y_t = k \mid a)] \)
- Finally

\[ NWGM[p(y_t = k \mid a)] = \frac{\prod_i \exp(n_{t,k,i})p(s_{t,i}=1|a)}{\sum_j \prod_i \exp(n_{t,j,i})p(s_{t,i}=1|a)} = \frac{\exp(\mathbb{E}_{p(s_t|a)}[n_{t,k}])}{\sum_j \exp(\mathbb{E}_{p(s_t|a)}[n_{t,j}])} \]

\[ \mathbb{E}[n_t] = L_o(Ey_{t-1} + L_h \mathbb{E}[h_t] + L_z \mathbb{E}[\hat{z}_t]) \]

How soft/hard attention works
How soft/hard attention works

A bird flying over a body of water.

Sample regions of attention

A variational lower bound of maximum likelihood

$$L_s = \sum_{s} p(s | a) \log p(y | s, a)$$

Computes the expected attention

$$\hat{z}_t = < p_1, p_2, p_3, p_4, p_5, p_6 >$$
<table>
<thead>
<tr>
<th>A</th>
<th>man</th>
<th>and</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>woman</td>
<td>playing</td>
</tr>
<tr>
<td>in</td>
<td>a</td>
<td>frisbee</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hard Attention
Soft Attention
The Good

A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background.

A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A **giraffe** standing in a forest with **trees** in the background.
And the Bad

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.
### Quantitative results

The attention-based model was further found to be highly interpretable, especially, compared to the simple encoder–decoder framework proposed in [25].

<table>
<thead>
<tr>
<th>Model</th>
<th>Human M1</th>
<th>Human M2</th>
<th>Automatic BLEU</th>
<th>Automatic CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.638</td>
<td>0.675</td>
<td>0.471</td>
<td>0.91</td>
</tr>
<tr>
<td>Google*</td>
<td>0.273</td>
<td>0.317</td>
<td>0.587</td>
<td>0.946</td>
</tr>
<tr>
<td>MSR*</td>
<td>0.268</td>
<td>0.322</td>
<td>0.567</td>
<td>0.925</td>
</tr>
<tr>
<td>Attention-based*</td>
<td>0.262</td>
<td>0.272</td>
<td>0.523</td>
<td>0.878</td>
</tr>
<tr>
<td>Captivator°</td>
<td>0.250</td>
<td>0.301</td>
<td>0.601</td>
<td>0.937</td>
</tr>
<tr>
<td>Berkeley LRCN°</td>
<td>0.246</td>
<td>0.268</td>
<td>0.534</td>
<td>0.891</td>
</tr>
</tbody>
</table>

M1: human preferred (or equal) the method over human annotation
M2: turing test

- Add soft attention to image captioning: +2 BLEU
- Add hard attention to image captioning: +4 BLEU
Video Description Generation

• Two encoders
  – Context set consists of per-frame context vectors, and attention mechanism that selects one of those vectors for each output symbol being decoded – capturing the global temporal structure across frames
  – 3-D conv-net that applies local filters across spatio-temporal dimensions working on motion statistics

• Both encoders are complementary
Memory Networks
Why attention?

• Long term memories - attending to memories
  – Dealing with gradient vanishing problem

• Exceeding limitations of a global representation
  – Attending/focusing to smaller parts of data
    ▪ patches in images
    ▪ words or phrases in sentences

• Decoupling representation from a problem
  – Different problems required different sizes of representations
    ▪ LSTM with longer sentences requires larger vectors

• Overcoming computational limits for visual data
  – Focusing only on the parts of images
  – Scalability independent of the size of images

• Adds some interpretability to the models (error inspection)
Attention on Memory Elements

- Recurrent networks cannot remember things for very long
  - The cortex only remember things for 20 seconds

- We need a “hippocampus” (a separate memory module)
  - LSTM [Hochreiter 1997], registers
  - Memory networks [Weston et al. 2014] (FAIR), associative memory
  - NTM [Graves et al. 2014], “tape”.
Recall: Long-Term Dependencies

- The RNN gradient is a product of Jacobian matrices, each associated with a step in the forward computation. To store information robustly in a finite-dimensional state, the dynamics must be contractive [Bengio et al 1994].

\[
L = L(s_T(s_{T-1}(\ldots s_{t+1}(s_t, \ldots))))
\]

\[
\frac{\partial L}{\partial s_t} = \frac{\partial L}{\partial s_T} \frac{\partial s_T}{\partial s_{T-1}} \ldots \frac{\partial s_{t+1}}{\partial s_t}
\]

- Problems:
  - sing. values of Jacobians > 1 → gradients explode
  - or sing. values < 1 → gradients shrink & vanish (Hochreiter 1991)
  - or random → variance grows exponentially

Storing bits robustly requires sing. values<1

Gradient clipping
Gated Recurrent Units & LSTM

• Create a path where gradients can flow for longer with self-loop

• Corresponds to an eigenvalue of Jacobian slightly less than 1

• LSTM is heavily used (Hochreiter & Schmidhuber 1997)

• GRU light-weight version (Cho et al 2014)
Delays & Hierarchies to Reach Farther

• Delays and multiple time scales, Elhihi & Bengio NIPS 1995, Koutnik et al ICML 2014

Large Memory Networks: Sparse Access Memory for Long-Term Dependencies

• A mental state stored in an external memory can stay for arbitrarily long durations, until evoked for read or write.

• Forgetting = vanishing gradient.

• Memory = larger state, avoiding the need for forgetting/vanishing.
Memory Networks

• Class of models that combine large memory with learning component that can read and write to it.

• Incorporates reasoning with attention over memory (RAM).

• Most ML has limited memory which is more-or-less all that’s needed for “low level” tasks e.g. object detection.
Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.

Where is the milk now?
Where is Joe?
Where was Joe before the office?
Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.

Where is the milk now? A: office
Where is Joe?
Where was Joe before the office?
Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office?
Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen
Scenario 2

1 Mr. Cropper was opposed to our hiring you.
2 Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him.
3 He says female teachers can't keep order.
4 He 's started in with a spite at you on general principles, and the boys know it.
5 They know he 'll back them up in secret, no matter what they do, just to prove his opinions.
6 Cropper is sly and slippery, and it is hard to corner him. '
7 `` Are the boys big? ''
8 queried Esther anxiously.
9 `` Yes.
10 Thirteen and fourteen and big for their age.
11 You can't whip 'em -- that is the trouble.
12 A man might, but they 'd twist you around their fingers.
13 You 'll have your hands full, I 'm afraid.
14 But maybe they 'll behave all right after all. ''
15 Mr. Baxter privately had no hope that they would, but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application.
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home.
18 He was a big, handsome man with a very suave, polite manner.
19 He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon.
20 Esther felt relieved.
Scenario 2

1 Mr. Cropper was opposed to our hiring you.
2 Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him.
3 He says female teachers can't keep order.
4 He 's started in with a spite at you on general principles, and the boys know it.
5 They know he 'll back them up in secret, no matter what they do, just to prove his opinions.
6 Cropper is sly and slippery, and it is hard to corner him. '
7 " Are the boys big? "
8 queried Esther anxiously.
9 " Yes."
10 Thirteen and fourteen and big for their age.
11 You can't whip 'em -- that is the trouble.
12 A man might, but they 'd twist you around their fingers.
13 You 'll have your hands full, I 'm afraid.
14 But maybe they 'll behave all right after all. '"
15 Mr. Baxter privately had no hope that they would, but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application.
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home.
18 He was a big, handsome man with a very suave, polite manner.
19 He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon.
20 Esther felt relieved.

Q: She thought that Mr. ______ had exaggerated matters a little.
Scenario 2

S: 1 Mr. Cropper was opposed to our hiring you.
2 Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him.
3 He says female teachers can't keep order.
4 He 's started in with a spite at you on general principles, and the boys know it.
5 They know he 'll back them up in secret, no matter what they do, just to prove his opinions.
6 Cropper is sly and slippery, and it is hard to corner him. '
7 `` Are the boys big? ''
8 queried Esther anxiously.
9 `` Yes.
10 Thirteen and fourteen and big for their age.
11 You can't whip 'em -- that is the trouble.
12 A man might, but they 'd twist you around their fingers.
13 You 'll have your hands full, I 'm afraid.
14 But maybe they 'll behave all right after all. ''
15 Mr. Baxter privately had no hope that they would, but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application.
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home.
18 He was a big, handsome man with a very suave, polite manner.
19 He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon.
20 Esther felt relieved.

Q: She thought that Mr. Baxter had exaggerated matters a little.
Scenario 3

*Shaolin Soccer* directed by *Stephen Chow*
*Shaolin Soccer* written by *Stephen Chow*
*Shaolin Soccer* starred actors *Stephen Chow*
*Shaolin Soccer* release year 2001
*Shaolin Soccer* has genre comedy
*Shaolin Soccer* has tags martial arts, kung fu soccer, *stephen chow*
*Kung Fu Hustle* directed by *Stephen Chow*
*Kung Fu Hustle* written by *Stephen Chow*
*Kung Fu Hustle* starred actors *Stephen Chow*
*Kung Fu Hustle* has genre comedy action
*Kung Fu Hustle* has imdb votes famous
*Kung Fu Hustle* has tags comedy, action, martial arts, kung fu, china, soccer, hong kong, *stephen chow*
*The God of Cookery* directed by *Stephen Chow*
*The God of Cookery* written by *Stephen Chow*
*The God of Cookery* starred actors *Stephen Chow*
*The God of Cookery* has tags hong kong *Stephen Chow*
*From Beijing with Love* directed by *Stephen Chow*
*From Beijing with Love* written by *Stephen Chow*
*From Beijing with Love* starred actors *Stephen Chow*, Anita Yuen

...<and more> ...
Scenario 3

**Shaolin Soccer** directed by Stephen Chow
**Shaolin Soccer** written by Stephen Chow
**Shaolin Soccer** starred actors Stephen Chow
**Shaolin Soccer** release year 2001
**Shaolin Soccer** has genre comedy
**Shaolin Soccer** has tags martial arts, kung fu soccer, stephen chow
**Kung Fu Hustle** directed by Stephen Chow
**Kung Fu Hustle** written by Stephen Chow
**Kung Fu Hustle** starred actors Stephen Chow
**Kung Fu Hustle** has genre comedy action
**Kung Fu Hustle** has imdb votes famous
**Kung Fu Hustle** has tags comedy, action, martial arts, kung fu, china, soccer, hong kong, stephen chow
**The God of Cookery** directed by Stephen Chow
**The God of Cookery** written by Stephen Chow
**The God of Cookery** starred actors Stephen Chow
**The God of Cookery** has tags hong kong
**From Beijing with Love** directed by Stephen Chow
**From Beijing with Love** written by Stephen Chow
**From Beijing with Love** starred actors Stephen Chow, Anita Yuen

... <and more> ...

Who wrote Kung Fu Hustle?
Scenario 3

Shaolin Soccer directed by Stephen Chow
Shaolin Soccer written by Stephen Chow
Shaolin Soccer starred actors Stephen Chow
Shaolin Soccer release year 2001
Shaolin Soccer has genre comedy
Shaolin Soccer has tags martial arts, kung fu soccer, stephen chow
Kung Fu Hustle directed by Stephen Chow
Kung Fu Hustle written by Stephen Chow
Kung Fu Hustle starred actors Stephen Chow
Kung Fu Hustle has genre comedy action
Kung Fu Hustle has imdb votes famous
Kung Fu Hustle has tags comedy, action, martial arts, kung fu, china, soccer, hong kong, stephen chow
The God of Cookery directed by Stephen Chow
The God of Cookery written by Stephen Chow
The God of Cookery starred actors Stephen Chow
The God of Cookery has tags hong kong Stephen Chow
From Beijing with Love directed by Stephen Chow
From Beijing with Love written by Stephen Chow
From Beijing with Love starred actors Stephen Chow, Anita Yuen
…<and more>…

I’m interested in watching a Stephen Chow movie other than Kung Fu Hustle. Can you suggest something?
What is required?

• Not all problems can be mapped to $y = f(x)$
• The model needs to remember external context
• Given an input, the model needs to know where to look for in the context
• It needs to know what to look for in the context
• It needs to know how to reason using this external context
• It needs to handle the potentially changing external context
What is a Memory Network?

*Original paper description of class of models*

MemNNs have four component networks (which may or may not have shared parameters):

- **I**: (input feature map) convert incoming data to the internal feature representation.
- **G**: (generalization) update memories given new input.
- **O**: produce new output (in feature representation space) given the memories.
- **R**: (response) convert output O into a response seen by the outside world.
Memory Networks - Some early publications

Memory Network Models

implemented models..

Supervision
(direct or
reward-based)

Output

Memory Module

Controller module

$m$

read

addressing

${\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_N}$

$q$

Input

Internal state
Vector (initially: query)

Memory vectors

Figure: Saina Sukhbaatar
Variants of the class...

Some options and extensions:

• **Representation of inputs and memories could use all kinds of encodings:** bag of words, RNN style reading at word or character level, etc.

• **Different possibilities for output module:** e.g. multi-class classifier or uses an RNN to output sentences.

• **If the memory is huge** (e.g. Wikipedia) we need to organize the memories. Solution: hash the memories to store in buckets (topics). Then, memory addressing and reading doesn’t operate on all memories.

• **If the memory is full,** there could be a way of removing one it thinks is most useless; i.e. it “forgets” somehow. That would require a scoring function of the utility of each memory..