

Text Summarization

Announcements

- Final exam: Dec. 21st, 1:10-4pm
- Class participation grades in courseworks:
 - 10% of grade
- AlphaGo documentary free screening. 5:30pm, Tuesday November 21, Roone Arledge Cinema, Lerner Hall. Register:

<https://www.eventbrite.com/e/movie-screening-alphago-tickets-39963928185>

HW4 comments

- Training can take a long time
 - Start early!
- To learn: LSTM, attention, ROUGE, +beam search
- Analysis
 - Select good, bad, random output
 - What are the problems?
 - What solution do you propose?
- Implementation of proposed solution will give additional extra credit

Today

Evaluation through user study

LSTM

Another neural network approach to
summarization

Evaluation

User Study: Objectives

- Does multi-document summarization help?
 - Do summaries help the user find information needed to perform a report writing task?
 - Do users use information from summaries in gathering their facts?
 - Do summaries increase user satisfaction with the online news system?
 - Do users create better quality reports with summaries?
 - How do full multi-document summaries compare with minimal 1-sentence summaries such as Google News?

User Study: Design

- Four parallel news systems
 - *Source documents only*; no summaries
 - *Minimal single sentence summaries* (Google News)
 - *Newsblaster summaries*
 - *Human summaries*
- All groups write reports given four scenarios
 - A task similar to analysts
 - Can only use Newsblaster for research
 - Time-restricted

User Study: Execution

- 4 scenarios
 - 4 event clusters each
 - 2 directly relevant, 2 peripherally relevant
 - Average 10 documents/cluster
- 45 participants
 - Balance between liberal arts, engineering
 - 138 reports
- Exit survey
 - Multiple-choice and open-ended questions
- Usage tracking
 - Each click logged, on or off-site

“Geneva” Prompt

- The conflict between Israel and the Palestinians has been difficult for government negotiators to settle. Most recently, implementation of the “road map for peace”, a diplomatic effort sponsored by
- Who participated in the negotiations that produced the Geneva Accord?
- Apart from direct participants, who supported the Geneva Accord preparations and how?
- What has the response been to the Geneva Accord by the Palestinians?

Measuring Effectiveness

- Score report content and compare across summary conditions
- Compare user satisfaction per summary condition
- Comparing where subjects took report content from

Summary Level	Pyramid Score
Level 1 (documents only)	0.3354
Level 2 (one sentence summary)	0.3757
Level 3 (System-X summary)	0.4269
Level 4 (Human summary)	0.4027

Table 2: Mean Pyramid Scores on Reports, Scenario 1 (Geneva Accords) excluded.

User Satisfaction

- More effective than a web search with Newsblaster
 - Not true with documents only or single-sentence summaries
- Easier to complete the task with summaries than with documents only
- Enough time with summaries than documents only
- Summaries helped most
 - 5% single sentence summaries
 - 24% Newsblaster summaries
 - 43% human summaries

User Study: Conclusions

- Summaries measurably improve a news browser's effectiveness for research
- Users are more satisfied with Newsblaster summaries are better than single-sentence summaries like those of Google News
- Users want search
 - Not included in evaluation

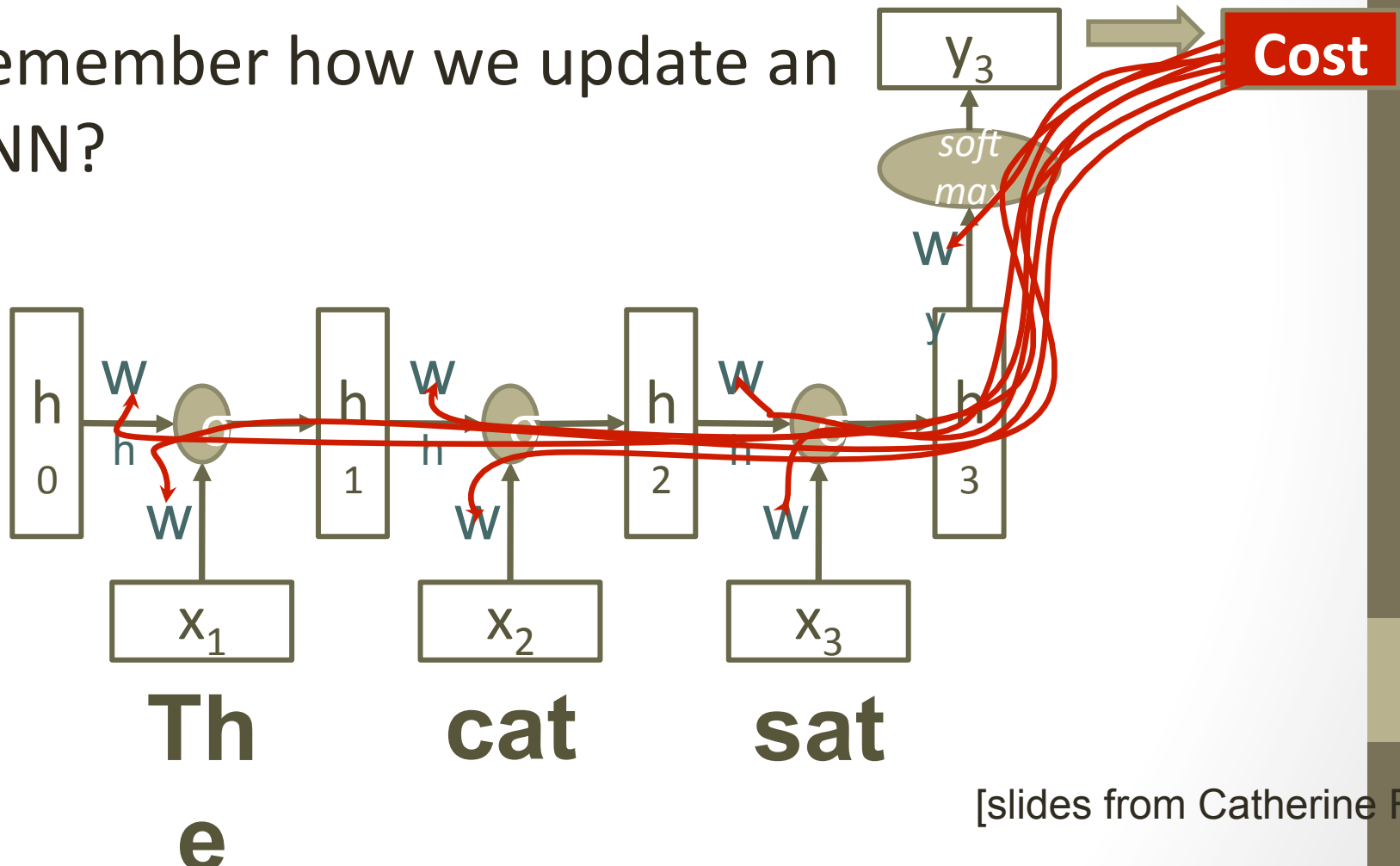
Questions (from Sparck Jones)

- Should we take the reader into account and how?
- Need more power than text extraction and more flexibility than fact extraction
- Evaluation: gold standard vs. user study?
Difficulty of evaluation?

Long Short-Term Memory Networks (LSTM)

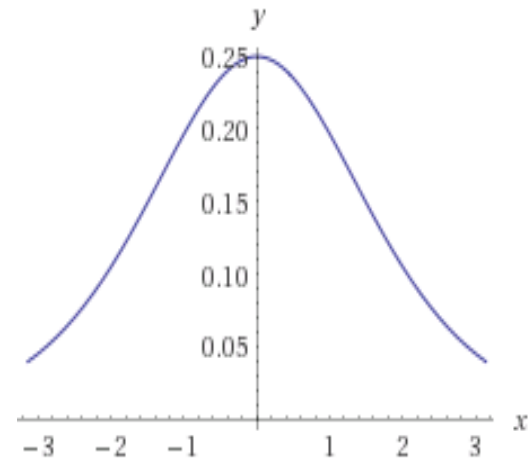
LSTM Motivation

Remember how we update an RNN?



The Vanishing Gradient Problem

- Deep neural networks use backpropagation.
- Back propagation uses the chain rule.
- The chain rule multiplies derivatives.
- Often these derivatives between 0 and 1.
- As the chain gets longer, products get smaller
 - until they disappear.



[slides from Catherine Finegan-Dollak]

Derivative of
sigmoid function

Wolfram

Or do they explode?


- With gradients larger than 1,
- you encounter the opposite problem
- with products becoming larger and larger
- as the chain becomes longer and longer,
- causing overlarge updates to parameters.
- This is the exploding gradient problem.

Vanishing/Exploding Gradients Are Bad.

- If we cannot backpropagate very far through the network, the network cannot learn long-term dependencies.

- My dog [chase/chases] squirrels. 

vs.

- My dog, whom I adopted in 2009, [chase/chases] squirrels. 

Gated Architectures

- RNN: at each state of the architecture, the entire memory state (h) is read and written
- Gate = binary vector $g \in \{0,1\}$
 - Controls access to n -dimensional vector $x \odot g$
- Consider $s' \leftarrow g \odot x + (1 - g) \odot (s)$
 - Reads entries from x specified by g
 - Copies remaining entries from s (or h as we've been labeling the hidden state)

$$\begin{array}{|c|} \hline 8 \\ \hline 11 \\ \hline 3 \\ \hline 7 \\ \hline 5 \\ \hline 15 \\ \hline \end{array} \leftarrow \begin{array}{|c|} \hline 0 \\ \hline 1 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline \end{array} \odot \begin{array}{|c|} \hline 10 \\ \hline 11 \\ \hline 12 \\ \hline 13 \\ \hline 14 \\ \hline 15 \\ \hline \end{array} + \begin{array}{|c|} \hline 1 \\ \hline 0 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 0 \\ \hline \end{array} \odot \begin{array}{|c|} \hline 8 \\ \hline 9 \\ \hline 3 \\ \hline 7 \\ \hline 5 \\ \hline 8 \\ \hline \end{array}$$

s' g x $(1-g)$ s

Example: gate copies from positions 2 and 5 in the input

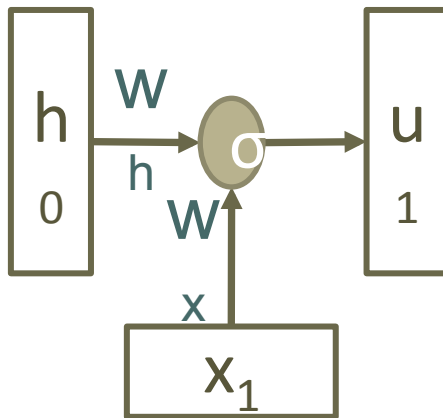
Remaining elements copied from memory

LSTM Solution

- Use memory cell to store information at each time step.
- Use “gates” to control the flow of information through the network.
 - Input gate: protect the current step from irrelevant inputs
 - Output gate: prevent the current step from passing irrelevant outputs to later steps
 - Forget gate: limit information passed from one cell to the next

Transforming RNN to LSTM

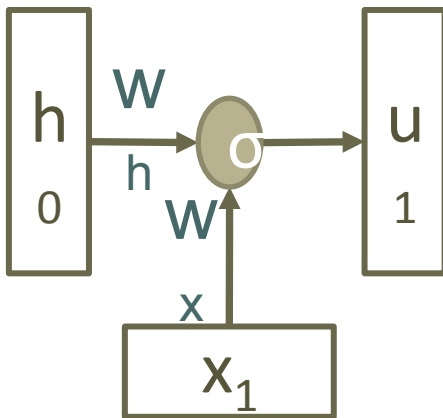
$$u_t = \sigma(W_h h_{t-1} + W_x x_t)$$



[slides from Catherine Finegan-Dollak]

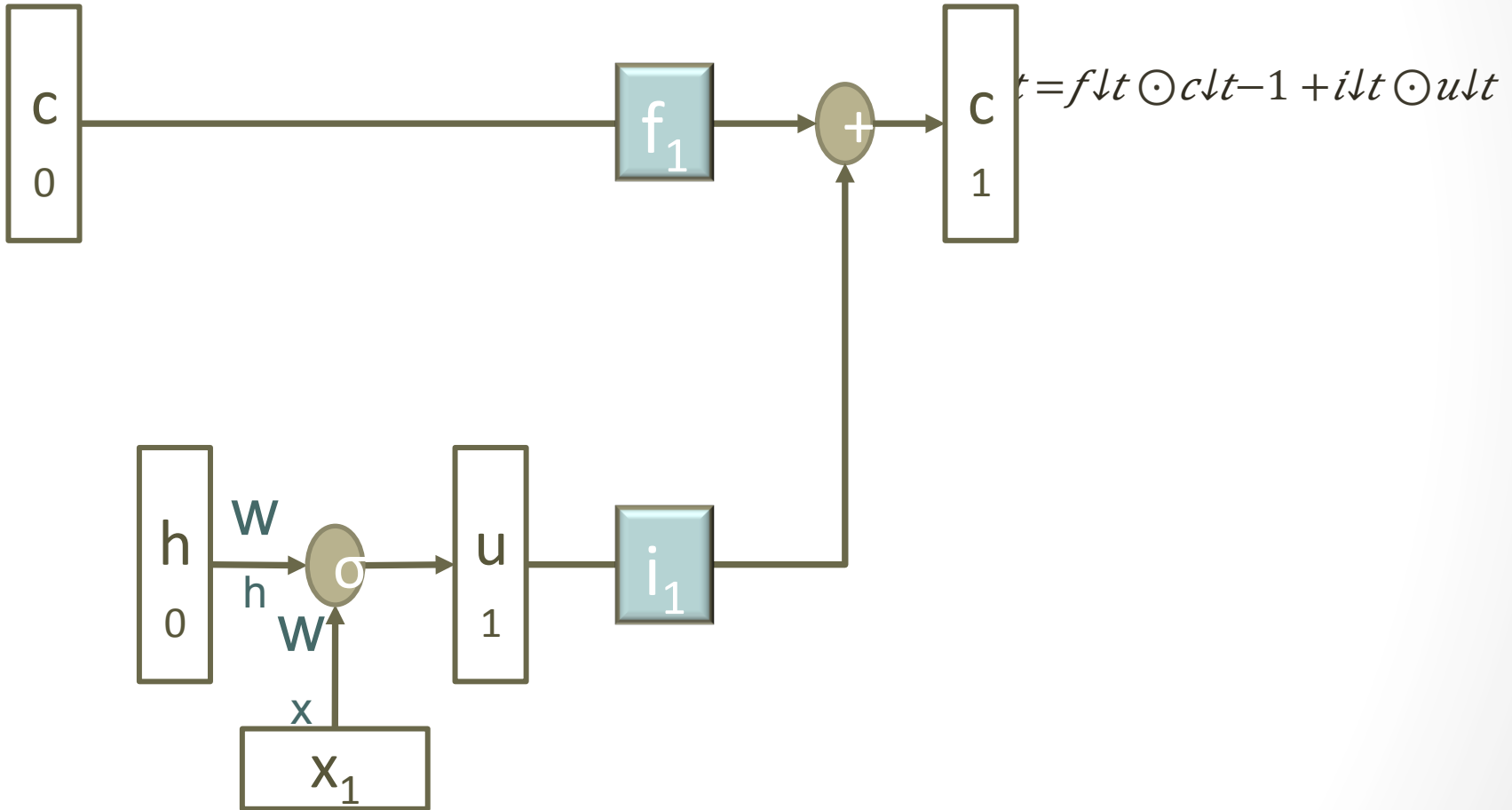
Transforming RNN to LSTM

C
0



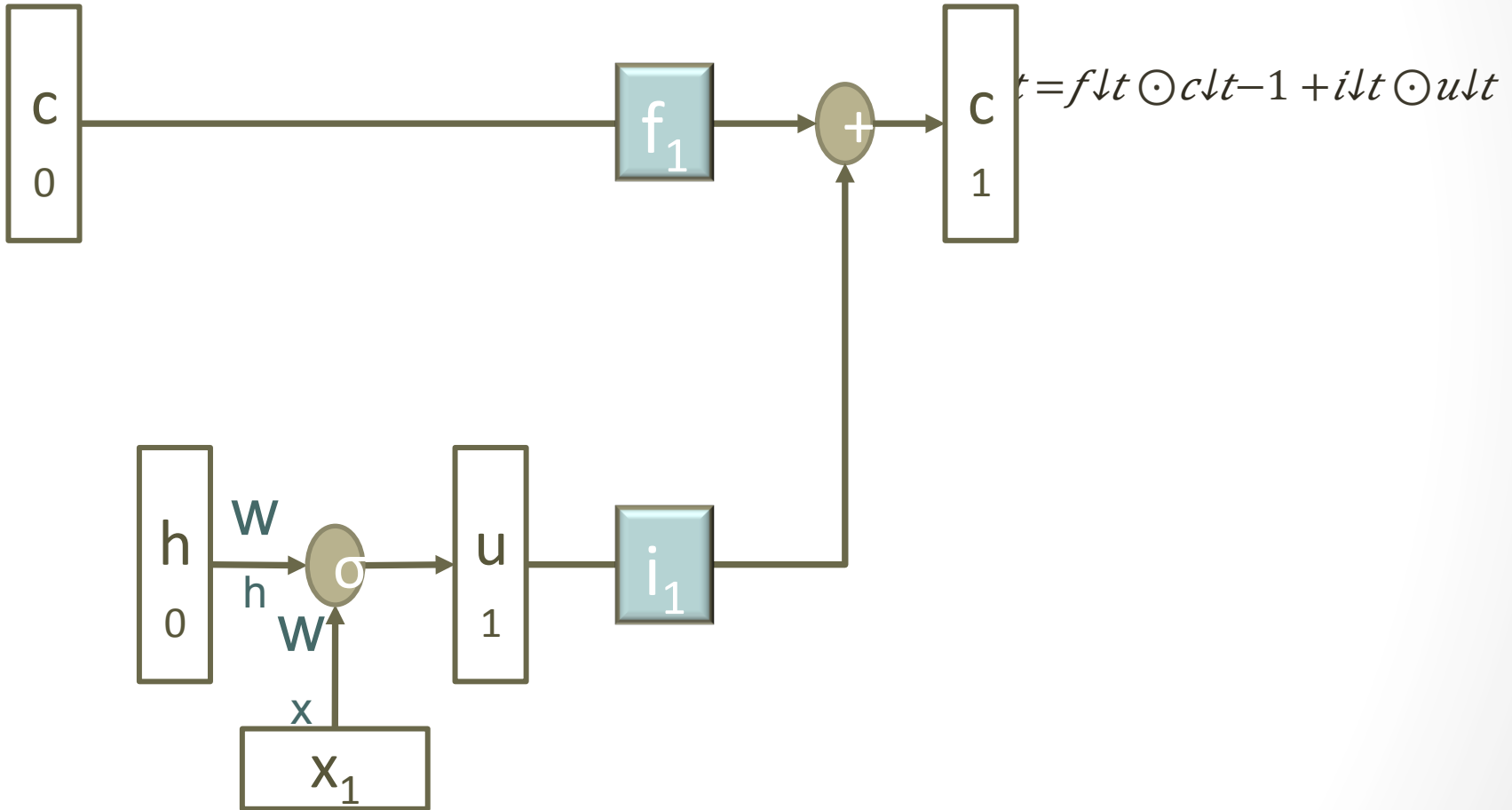
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM

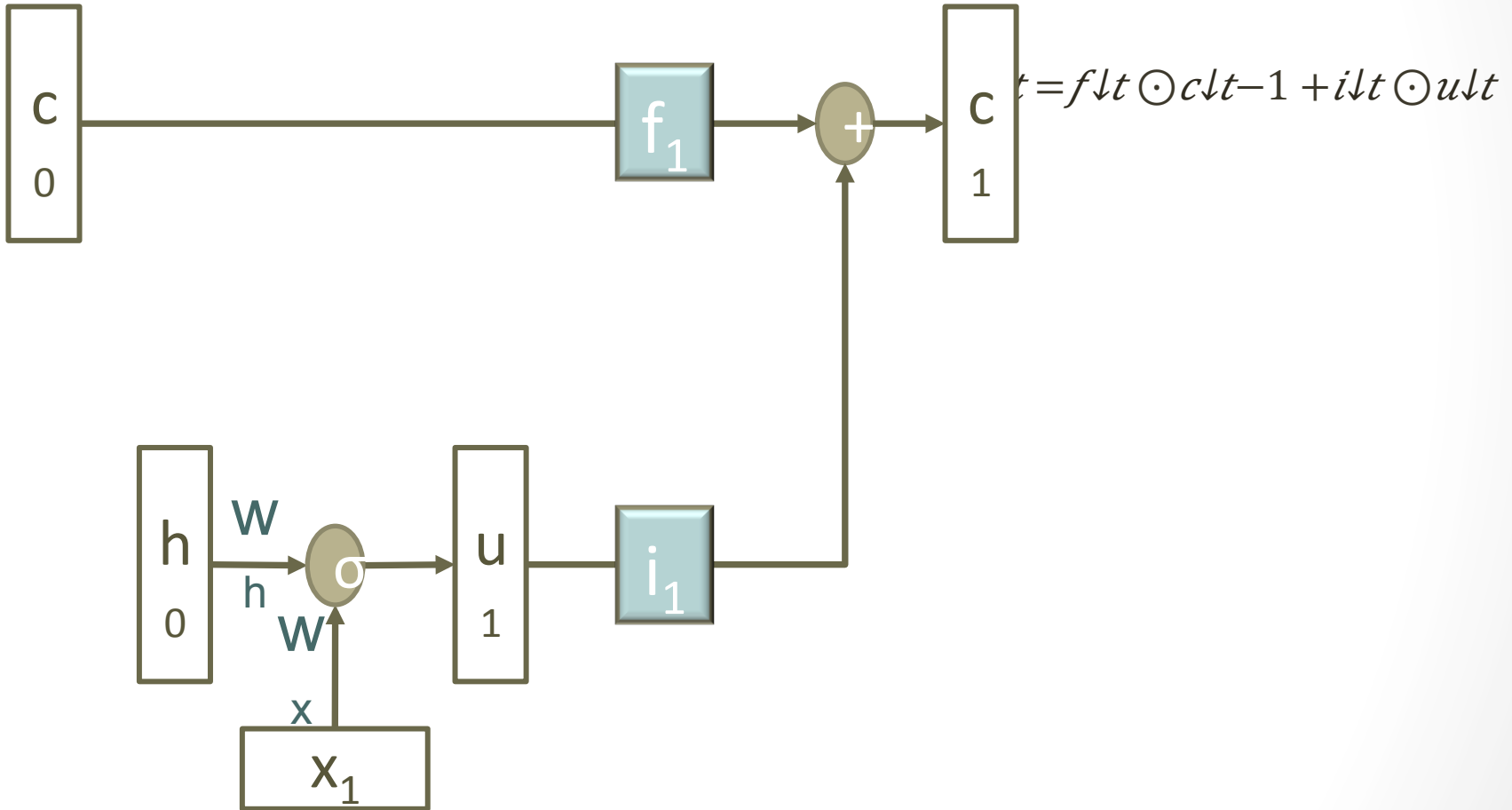


[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM

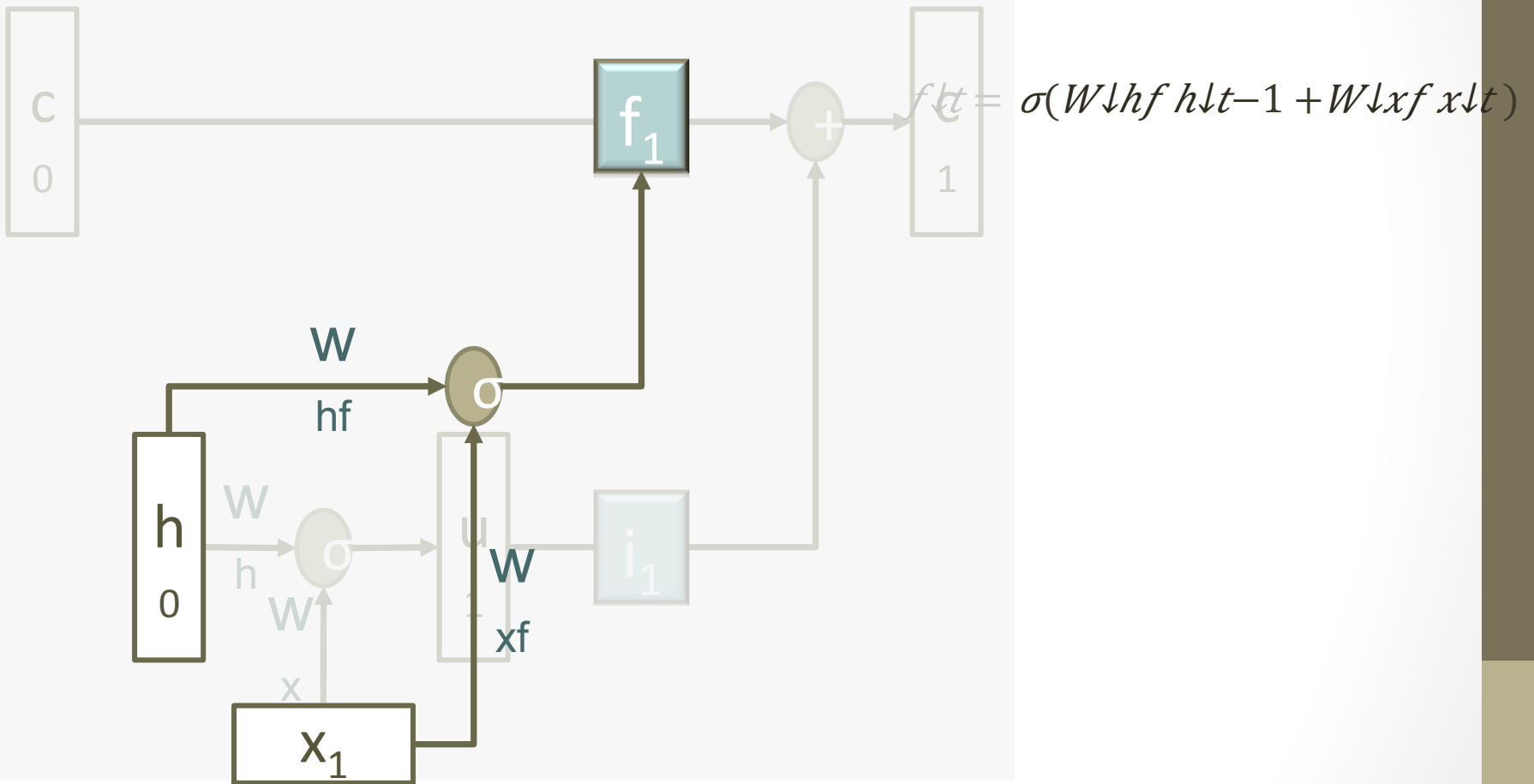


Transforming RNN to LSTM



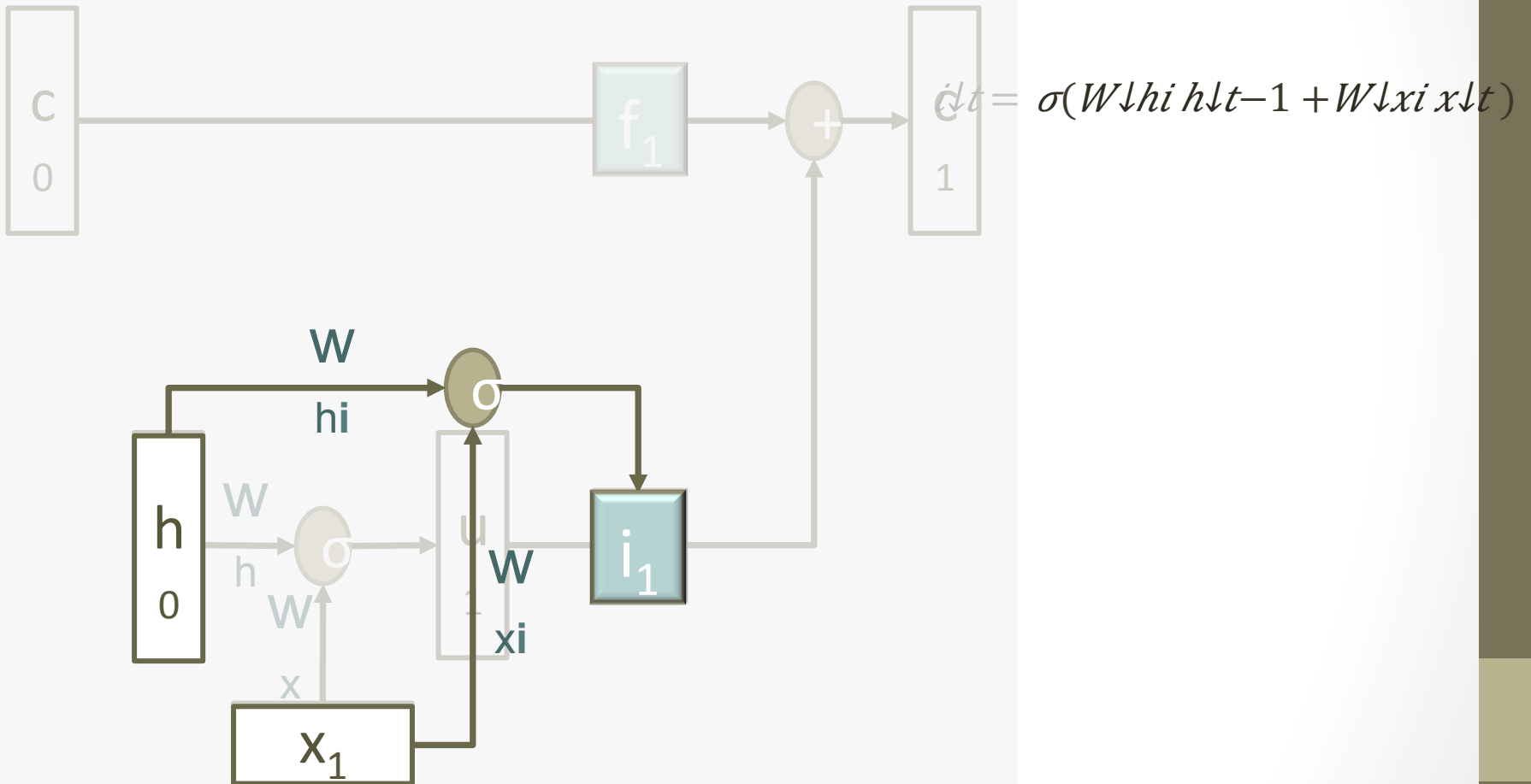
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM



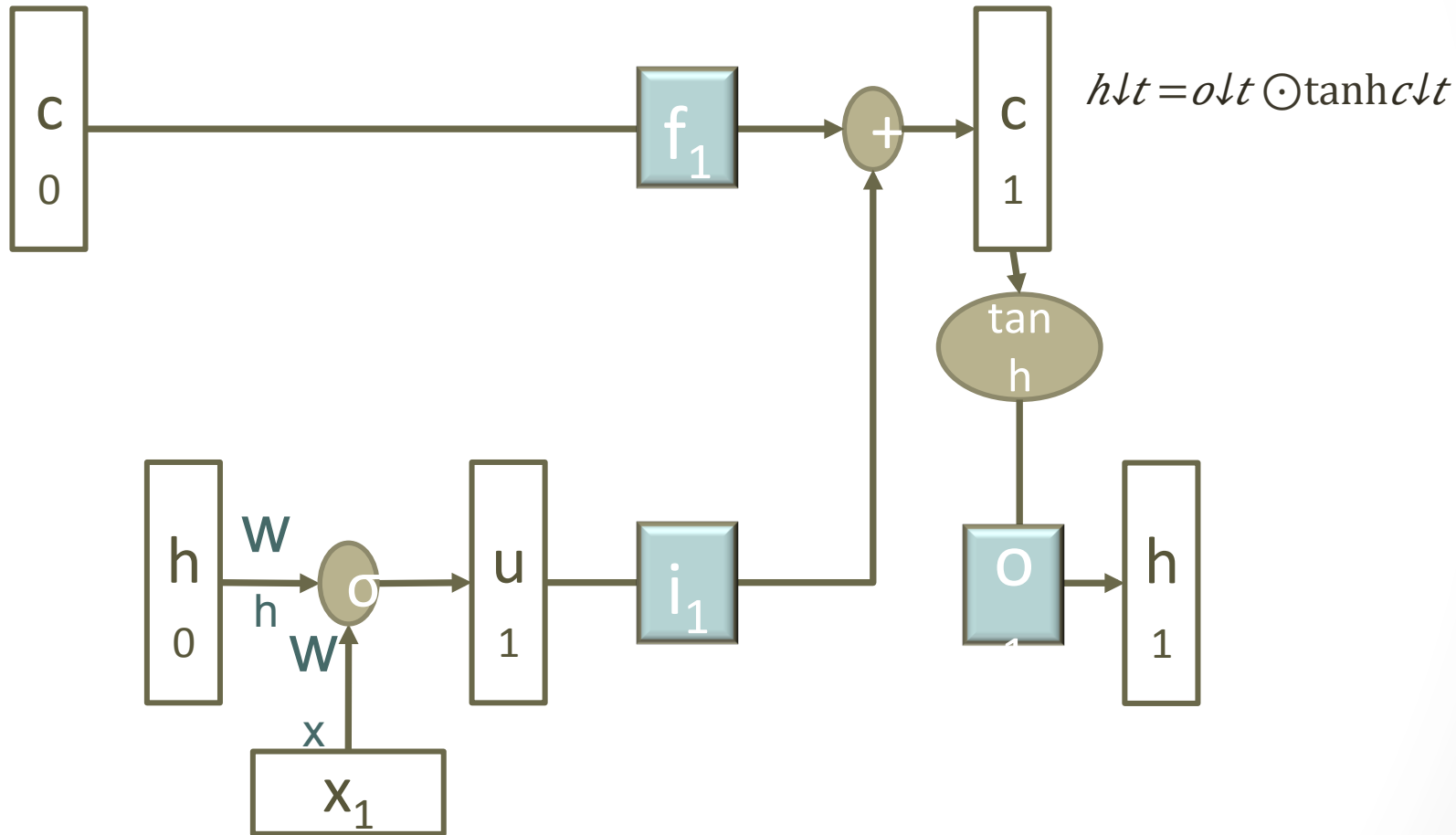
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM



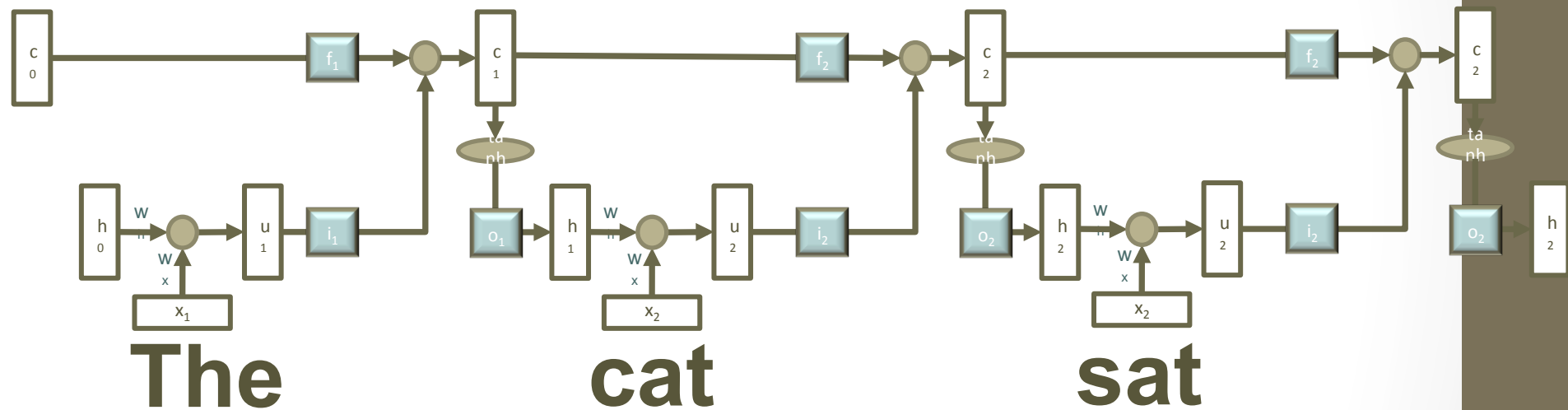
[slides from Catherine Finegan-Dollak]

Transforming RNN to LSTM



[slides from Catherine Finegan-Dollak]

LSTM for Sequences



LSTM Applications

Handwriting generation

<http://www.cs.toronto.edu/~graves/handwriting.html>

- Language identification (Gonzalez-Dominguez et al., 2014)
- Paraphrase detection (Cheng & Kartsaklis, 2015)
- Speech recognition (Graves, Abdel-Rahman, & Hinton, 2013)
- Handwriting recognition (Graves & Schmidhuber, 2009)
- Music composition (Eck & Schmidhuber, 2002) and lyric generation (Potash, Romanov, & Rumshisky, 2015)
- Robot control (Mayer et al., 2008)
- Natural language generation (Wen et al. 2015) (best paper at EMNLP)
- Named entity recognition (Hammerton, 2003)

Another neural summarization approach

- Extractive summarization of news
 - Single document summarization
- Data source: Daily News
 - Bulleted highlights of each article
- Neural Summarization by Extracting Sentences and Words
 - Cheng and Lapata, Edinburgh

Example from Daily News

AFL star blames vomiting cat for speeding

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia \$824 for exceeding the speed limit by more than 30km/h.

He lost four demerit points, instead of seven, because of his significant training commitments.

- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined \$824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

Figure 1: DailyMail news article with highlights. Underlined sentences bear label 1, and 0 otherwise.

Example from Daily News

AFL star blames vomiting cat for speeding

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia \$824 for exceeding the speed limit by more than 30km/h.


He lost four demerit points, instead of seven, because of his significant training commitments.

- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined \$824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

Figure 1: DailyMail news article with highlights. Underlined sentences bear label 1, and 0 otherwise.

Paraphrasing
Compression
Fusion

see examples of paraphrasing, fusion, compr



paraphrasing


fusion

compression



Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app



Example from Daily News

AFL star blames vomiting cat for speeding

Adelaide Crows defender Daniel Talia has kept his driving license, telling a court he was speeding 36km over the limit because he was distracted by his sick cat.

The 22-year-old AFL star, who drove 96km/h in a 60km/h road works zone on the South Eastern expressway in February, said he didn't see the reduced speed sign because he was so distracted by his cat vomiting violently in the back seat of his car.

In the Adelaide magistrates court on Wednesday, Magistrate Bob Harrap fined Talia \$824 for exceeding the speed limit by more than 30km/h.

He lost four demerit points, instead of seven, because of his significant training commitments.


- *Adelaide Crows defender Daniel Talia admits to speeding but says he didn't see road signs because his cat was vomiting in his car.*
- *22-year-old Talia was fined \$824 and four demerit points, instead of seven, because of his 'significant' training commitments.*

Figure 1: DailyMail news article with highlights. Underlined sentences bear label 1, and 0 otherwise.


Paraphrasing
Compression
Fusion

Two Tasks

- Input: Document $D: \{s_1, \dots, s_m\}$ consisting of words w_1, \dots, w_n
- Sentence extraction
 - Select a subset of j sentences, $j < m$
 - Score each sentence and predict label $y_L \in \{0, 1\}$
 - Objective: Maximize all sentence labels given D and weights θ
- Word extraction
 - Find a subset of words in D and their optimal ordering
 - Language generation task with output vocabulary restricted to input D vocabulary
 - Objective: Maximize the likelihood of generated sentences, further decomposed by considering conditional dependencies among their words



word extraction different from your homework
(neural abstractive summarization)?



Start the presentation to activate live content



If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Training Data

- Sentence extraction
 - Highlights are abstracts
 - Find the s in D that most closely matches a highlight sentence
 - Positive, unigram and bigram matches, #entities
 - 200K document/summary pairs, summary size = 30% document
- Word extraction
 - Retain highlights with all words from D
 - Find neighbors of words not in D and substitute
 - 170K document/summary pairs

Neural Summarization Architecture

- Hierarchical document reader
 - Derive meaning representation of document from its constituent sentences
- Attention based hierarchical content extractor
- Encoder-decoder architecture

Document Reader

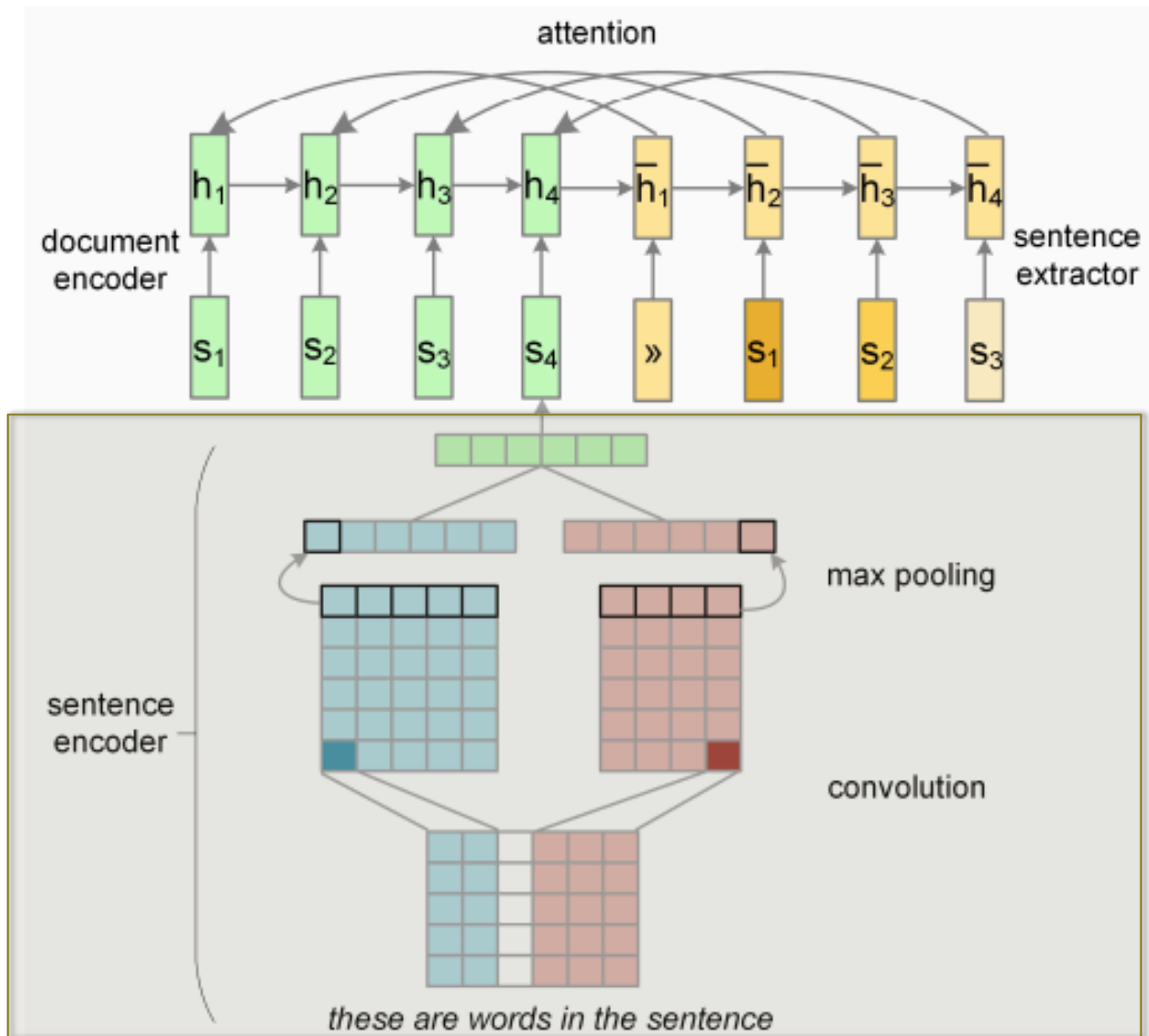
- CNN sentence encoder
 - Useful for sentence classification
 - Easy to train
- LSTM document encoder
 - Avoids vanishing gradients

CNN

$$\mathbf{f}_j^i = \tanh(\mathbf{W}_{j:j+c-1} \otimes \mathbf{K} + b)$$

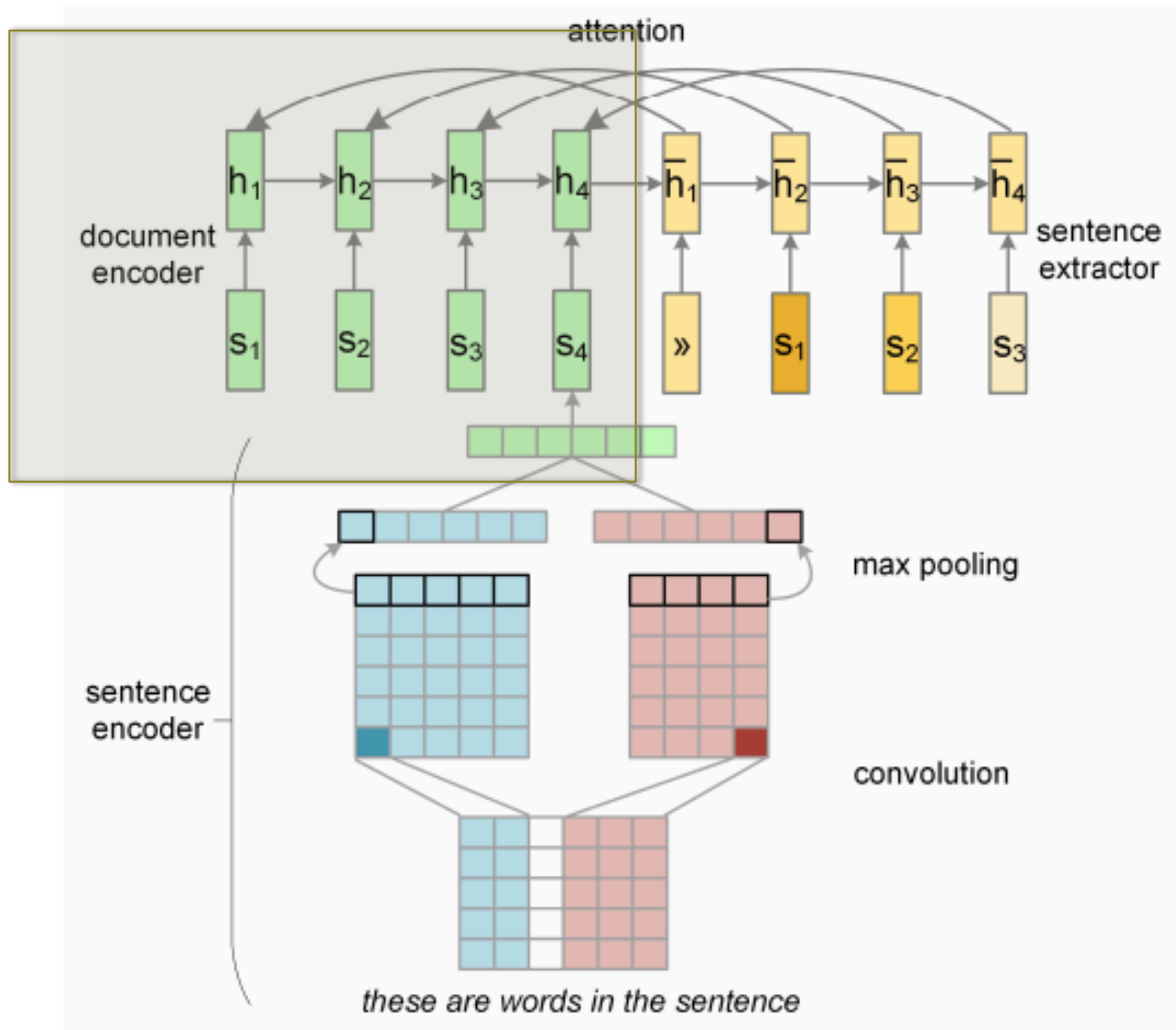
- Where $W \in \mathbb{R}^{n \times d}$ and $d =$ word embedding dimension, $n =$ #words in sentence
- K a kernel of width c , b the bias
- $f_j^i =$ the j th item in the i th feature map f^i
- Perform max pooling over time to obtain a single feature to represent the sentence

$$\mathbf{s}_{i,K} = \max_j \mathbf{f}_j^i$$



Recurrent document encoder

- LSTM to compose a sequence of sentence vectors into a document vector
- The hidden states of the LSTM = a list of partial representations
 - Each focuses on the corresponding input sentence given previous content
- Altogether constitute document representation



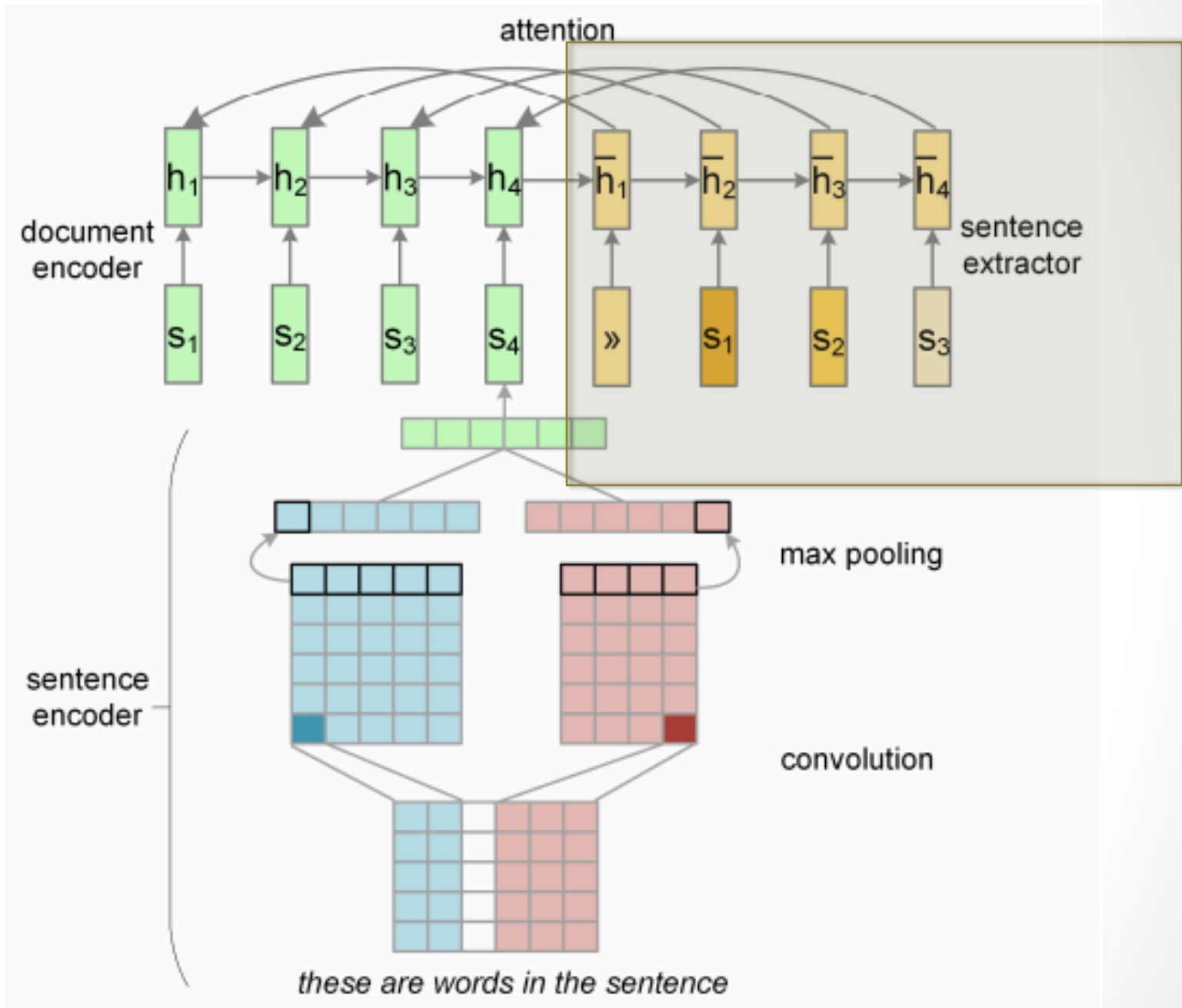
Sentence Extractor

- Applies attention to directly extract sentences after reading them

$$\bar{\mathbf{h}}_t = \text{LSTM}(p_{t-1}\mathbf{s}_{t-1}, \bar{\mathbf{h}}_{t-1})$$

$$p(y_L(t) = 1|D) = \sigma(\text{MLP}(\bar{\mathbf{h}}_t : \mathbf{h}_t))$$

- $\bar{\mathbf{h}}$ extractor hidden state, \mathbf{h} encoder hidden state
 - Attends to relation between extractor and encoder hidden state
- MLP takes as input concatenated $\bar{\mathbf{h}}$ and \mathbf{h}
- P_{t-1} degree to which extractor believes previous sentence should be extracted



Word Extractor

- Instead of extracting sentence, extracts next word
- Uses hierarchical attention to attend to sentence and word within sentence
- Output vocabulary restricted to input sentence
- -> conditional language model with vocabulary constraint

Datasets

- Daily Mail
 - 200K training
 - 500 test
- DUC 2002
 - 567 documents with 2 summaries each

Results

DUC 2002	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	43.6	21.0	40.2
LREG	43.8	20.7	40.3
ILP	45.4	21.3	42.8
NN-ABS	15.8	5.2	13.8
TGRAPH	48.1	24.3	—
URANK	48.5	21.5	—
NN-SE	47.4	23.0	43.5
NN-WE	27.0	7.9	22.8

DailyMail	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	20.4	7.7	11.4
LREG	18.5	6.9	10.2
NN-ABS	7.8	1.7	7.1
NN-SE	21.2	8.3	12.0
NN-WE	15.7	6.4	9.8

Table 1: ROUGE evaluation (%) on the DUC-2002 and 500 DailyMail samples.

Next

Machine Translation

Happy Thanksgiving!!

Have a good break!