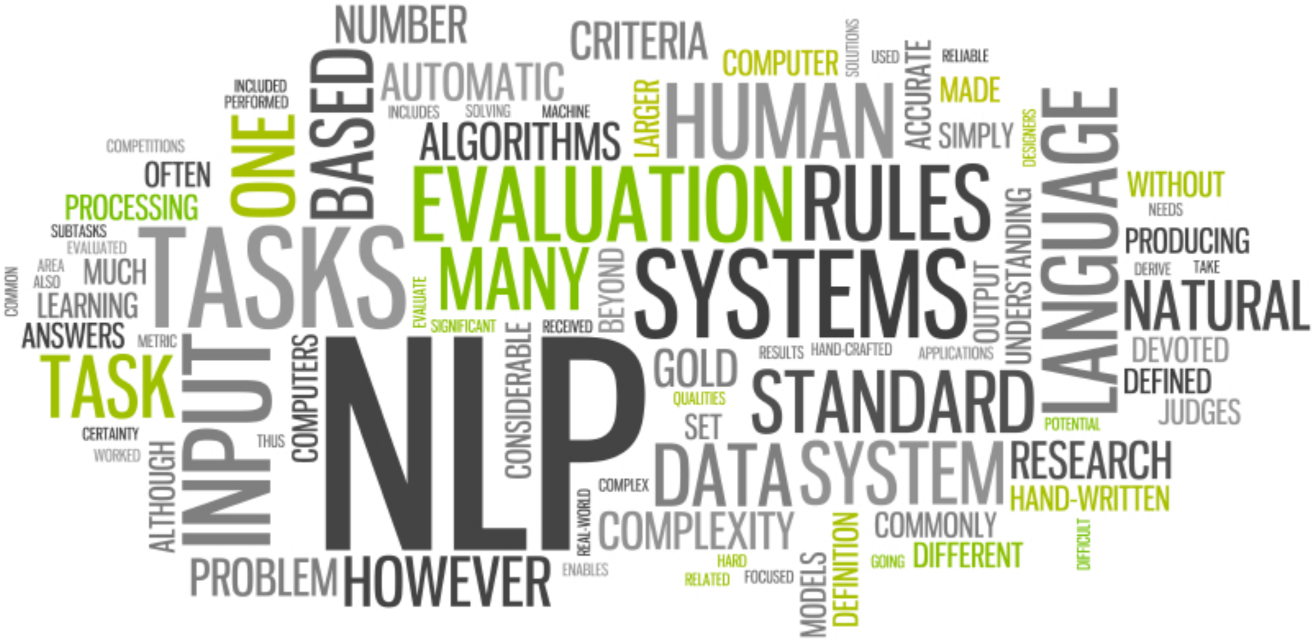
Natural Language Processing and Deep Learning Basics

He He

gluon-nlp.mxnet.io

Challenges and Tasks



Challenges in Natural Language Processing

Challenges in Natural Language Processing

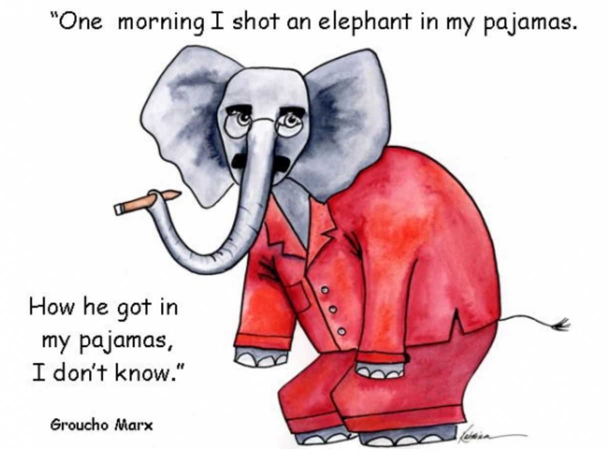
- Ambiguity

Challenges in Natural Language Processing

- Ambiguity
    - Time **flies** like an arrow; fruit **flies** like a banana.

Challenges in Natural Language Processing

- Ambiguity
  - Time **flies** like an arrow; fruit **flies** like a banana.
  - I shot an elephant in my pajama.



Challenges in Natural Language Processing

- Symbolic and discrete

small, tiny,
large, round

Challenges in Natural Language Processing

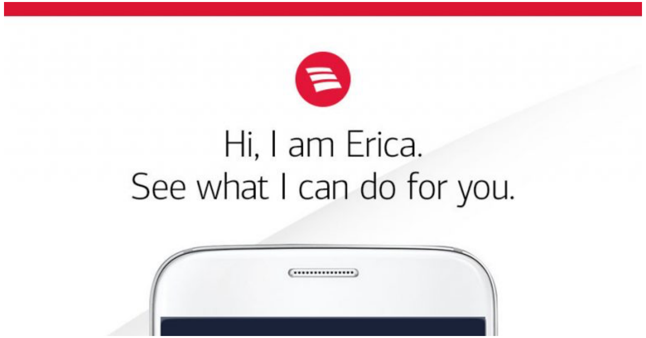
Challenges in Natural Language Processing

- Compositional
  - Letter -> word -> phrase -> sentence -> paragraph
  - Numerous ways to form a sentence

Challenges in Natural Language Processing

- Compositional
  - Letter -> word -> phrase -> sentence -> paragraph
  - Numerous ways to form a sentence
- Data sparsity

There are over 2,000 different
ways to ask Erica to move money.



NLP Tasks

NLP Tasks

Input	Output	Task

NLP Tasks

Input	Output	Task
Sentence(s)	Category	Sentiment analysis Document classification

NLP Tasks

Input	Output	Task
Sentence(s)	Category	Sentiment analysis Document classification
Sentence	Sequence of labels	Parts-of-speech tagging Named entity recognition

NLP Tasks

Input	Output	Task
Sentence(s)	Category	Sentiment analysis Document classification
Sentence	Sequence of labels	Parts-of-speech tagging Named entity recognition
Sentence	Tree	Parsing Relation extraction

NLP Tasks

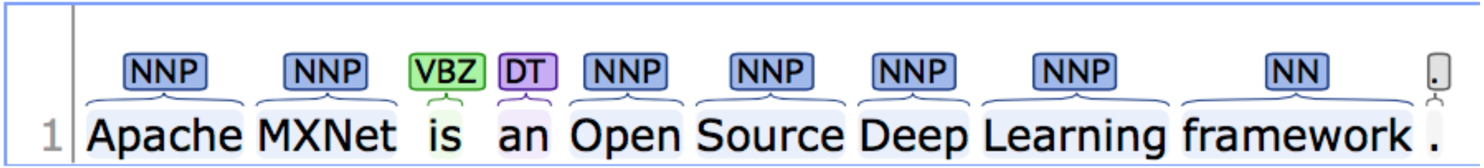
Input	Output	Task
Sentence(s)	Category	Sentiment analysis Document classification
Sentence	Sequence of labels	Parts-of-speech tagging Named entity recognition
Sentence	Tree	Parsing Relation extraction
Sentence	Sentence	Machine translation Question answering

Classification

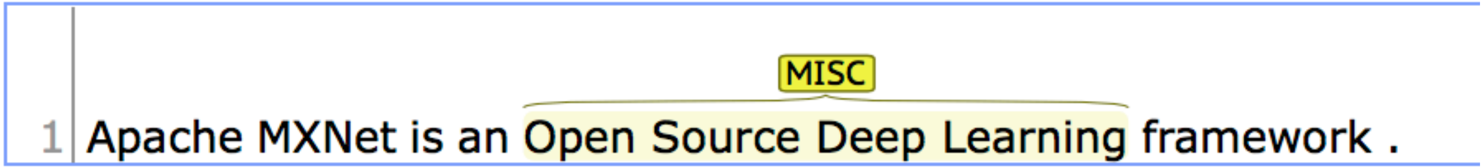
- Sentiment
  - The movie was horrible. ☹️
- Topic
  - Kevin Durant reportedly made the Warriors 'begrudgingly' give up… 🏀
- Entailment
  - The woman bought some flowers from the market.
  - The woman got flowers. 👍

Sequence Labeling

Part-of-Speech:



Named Entity Recognition:



Reading Comprehension

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title…

Which NFL team represented the AFC at Super Bowl 50?

Generation

- Machine translation
  - deep learning has transformed NLP
  - 深度学习改变了NLP
- Story generation
  - **Prompt:** The Mage, the Warrior, and the Priest

    Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

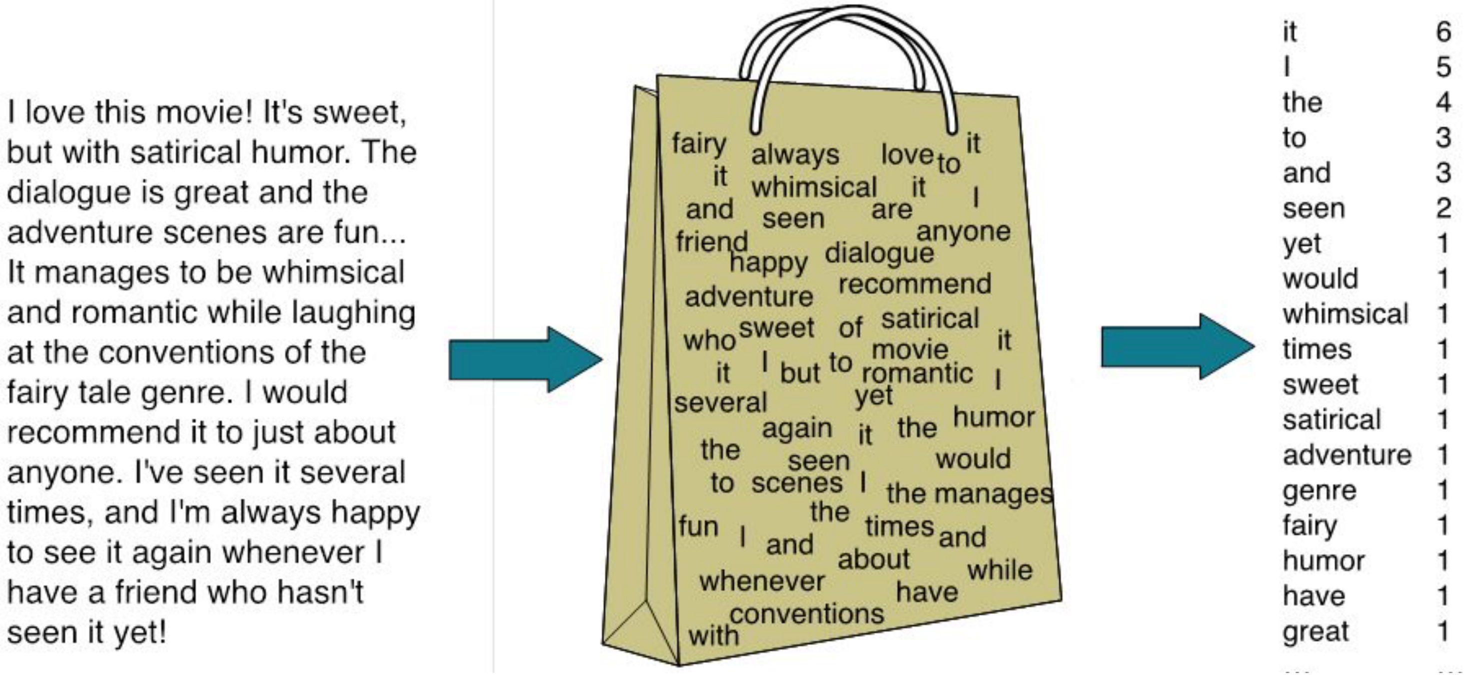
Features and Representations

Features of Words

- Surface form: *organizing*
- To reduce inflection
  - Lemma: *organize*
  - Stem: *organiz*
- Prefix / Suffix
- Length

Features of Sentences

Bag-of-words



Symbolic representation

Symbolic representation

- Words as discrete symbols
  - Vocab = {car, truck, cat, dog, house}
  - car = [1 0 0 0 0]
  - truck = [0 1 0 0 0]

Symbolic representation

- Words as discrete symbols
  - Vocab = {car, truck, cat, dog, house}
  - car = [1 0 0 0 0]
  - truck = [0 1 0 0 0]

Othogonal!

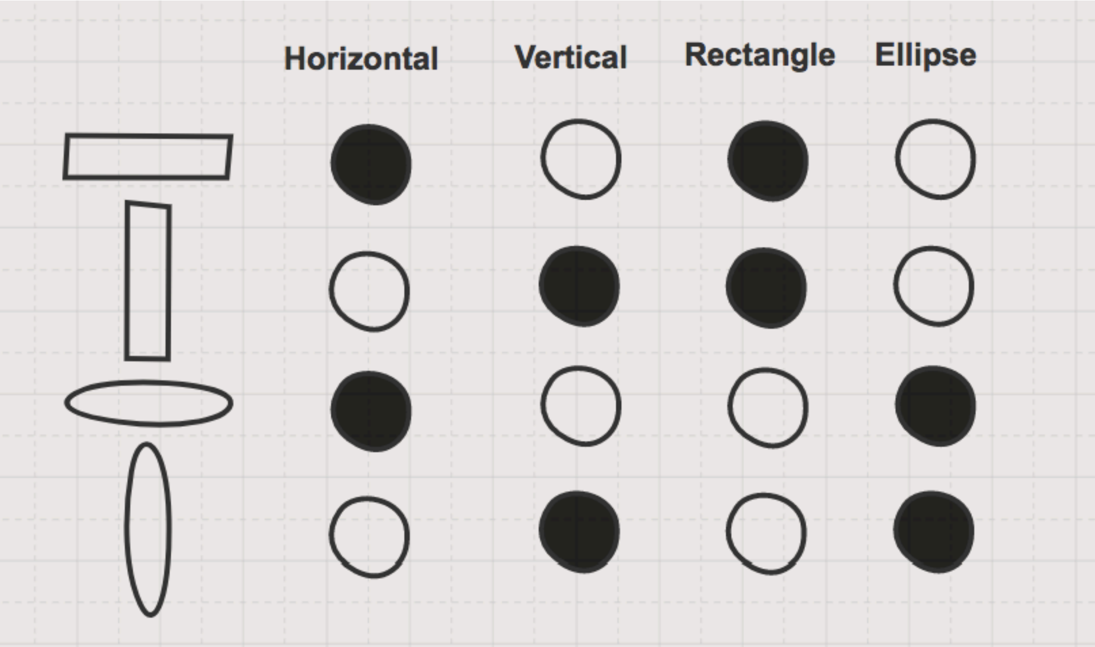
Symbolic representation

- Words as discrete symbols
    - Vocab = {car, truck, cat, dog, house}
    - car = [1 0 0 0 0]
    - truck = [0 1 0 0 0]

Othogonal!

- Problems
    - No semantic similarity
    - Sparse, high-dimensional (for large vocabulary)

Distributed Representation



(image courtesy of Garrett Hoffman)

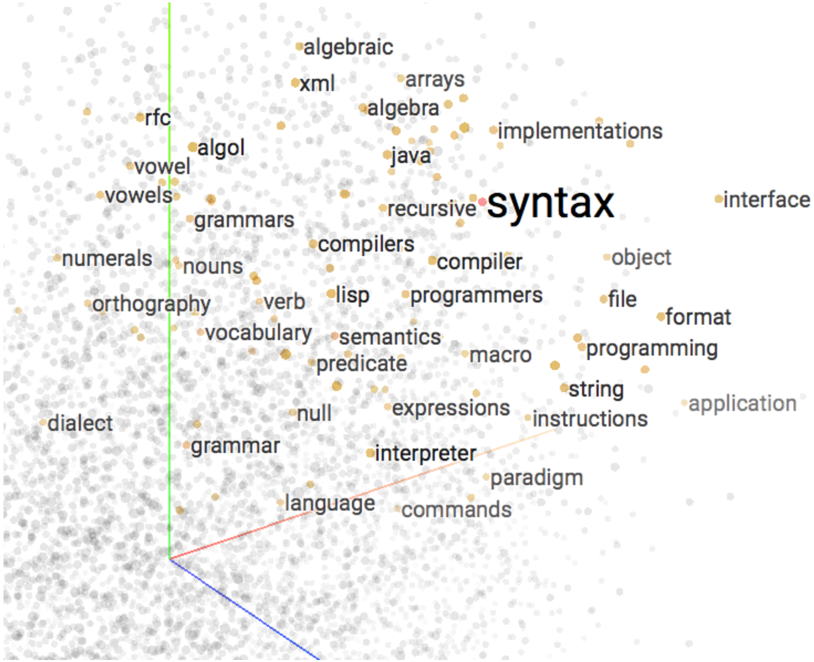
Represent Words by Context

“You shall know a word by the company it keeps” (J. R. Firth 1957: 11)

… full size **cars**, SUVs, and luxury vehicles …
… new or used small or full-sized **truck …**
A **truck** or lorry is a motor vehicle …
… new & used **cars**, compare models …

Word Vectors / Embeddings

$$\text{syntax} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$



Architectures for text

Modeling Language 101

- Assign probability to a sequence of words

Predict next word

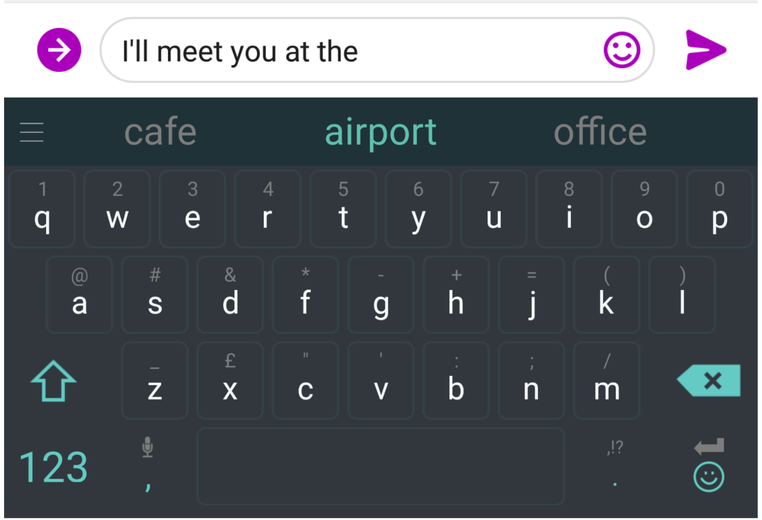
$$p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_1, \ldots, w_{t-1})$$

Modeling Language 101

- Assign probability to a sequence of words

Predict next word

$$p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_1, \ldots, w_{t-1})$$



gluon-nlp.mxnet.io

Modeling Language 101

- Assign probability to a sequence of words

$$
\begin{aligned}
& p(\text{Statistics}, \text{is}, \text{fun}, .\,) \\
= \; & p(\text{Statistics})p(\text{is} \mid \text{Statistics})p(\text{fun} \mid \text{Statistics}, \text{is})p(\,.\mid \text{Statistics}, \text{is}, \text{fun})
\end{aligned}
$$

- Estimating it

$$\hat{p}(\text{is} \mid \text{Statistics}) = \frac{\text{count}(\text{Statistics is})}{\text{count}(\text{Statistics})}$$

Problems

Problems

- Sparsity
  - count(Statistics if fun and oops) = 0
  - Need smoothing

Problems

- Sparsity
    - count(Statistics if fun and oops) = 0
    - Need smoothing
- Storage
    - Need to store counts of all ngrams in training data

Problems

- Sparsity
  - count(Statistics if fun and oops) = 0
  - Need smoothing
- Storage
  - Need to store counts of all ngrams in training data

Both don’t scale well with n

Solution

Solution

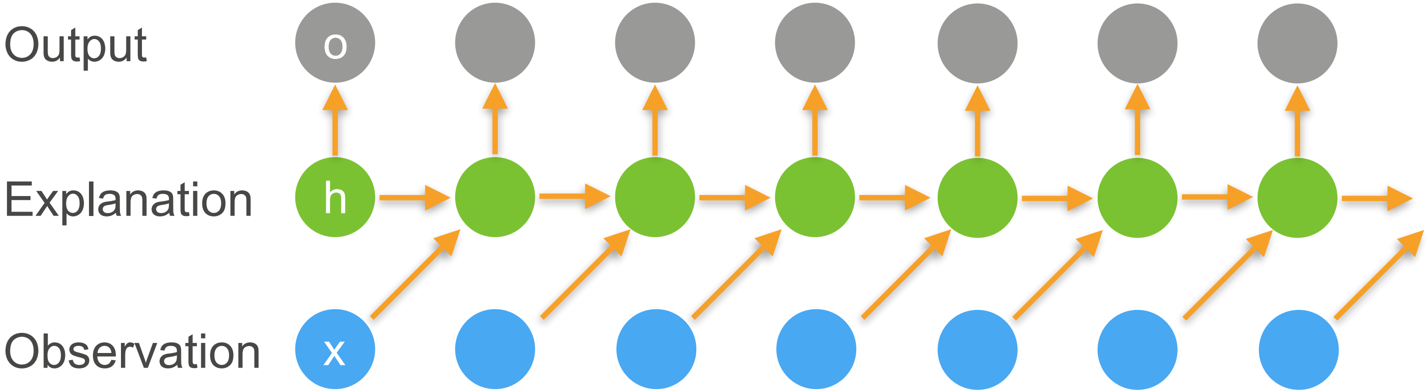
- $p(w_t \mid w_1, \ldots, w_{t-1})$: classifier with word embeddings as input

Solution

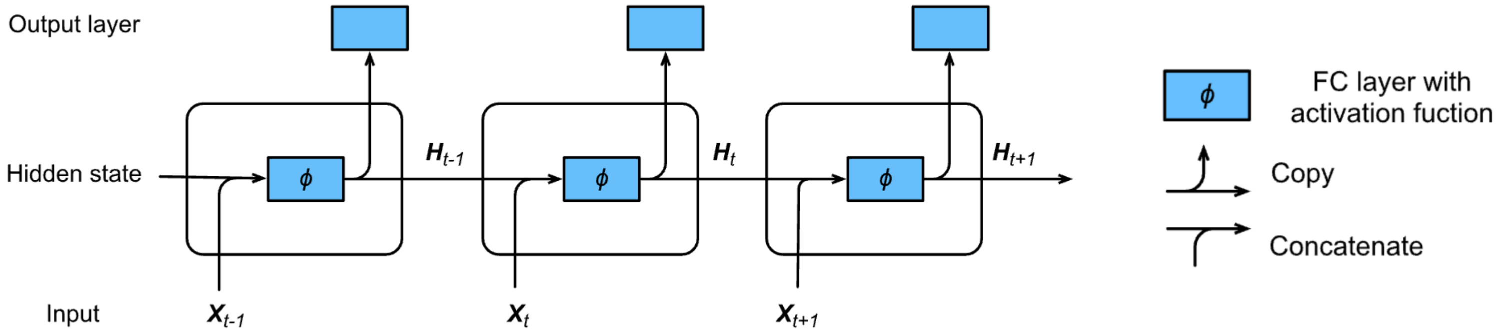
- $p(w_t \mid w_1, \ldots, w_{t-1})$: classifier with word embeddings as input
- Left-to-right text processing: recurrently update the input representation

Solution

- $p(w_t \mid w_1, \ldots, w_{t-1})$: classifier with word embeddings as input
- Left-to-right text processing: recurrently update the input representation

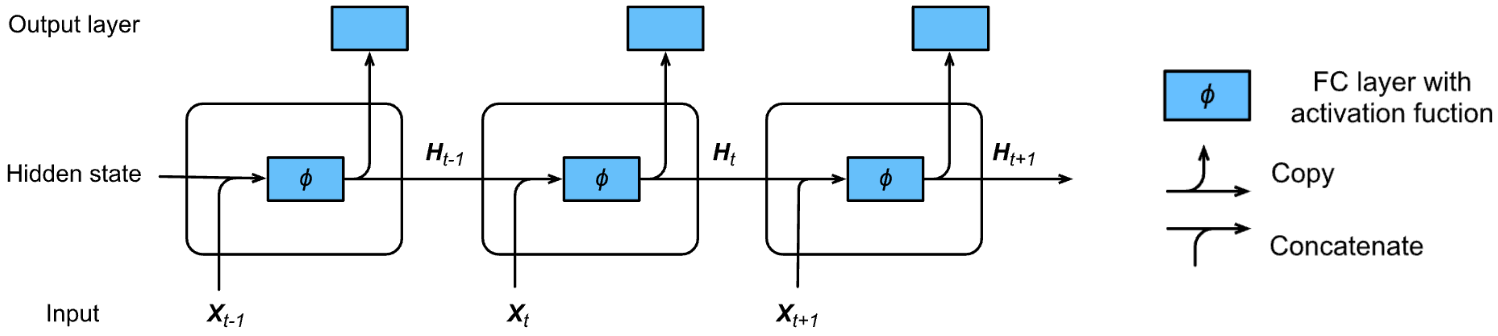


Recurrent Neural Networks



http://amlc-d2l.corp.amazon.com/

Recurrent Neural Networks



- Hidden State update

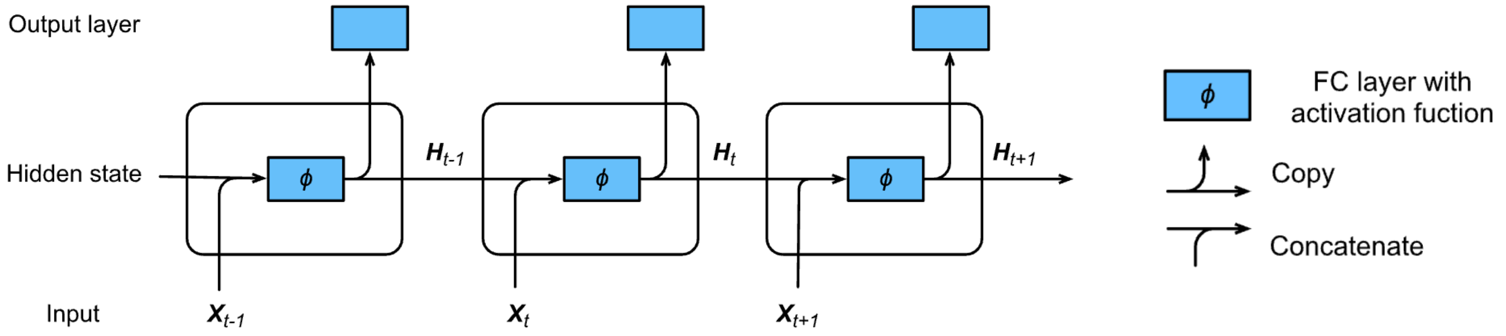
$$\mathbf{H}_t = \phi(\mathbf{W}_{hh}\mathbf{H}_{t-1} + \mathbf{W}_{hx}\mathbf{X}_{t-1} + \mathbf{b}_h)$$

- Observation update

$$\mathbf{o}_t = \mathbf{W}_{ho}\mathbf{H}_t + \mathbf{b}_o$$

http://amlc-d2l.corp.amazon.com/

Recurrent Neural Networks



- Hidden State update

$$\mathbf{H}_t = \phi(\boxed{\mathbf{W}_{hh}\mathbf{H}_{t-1}} + \mathbf{W}_{hx}\mathbf{X}_{t-1} + \mathbf{b}_h)$$

- Observation update

$$\mathbf{o}_t = \mathbf{W}_{ho}\mathbf{H}_t + \mathbf{b}_o$$

- Compare to MLP

$$\mathbf{H}_t = \phi(\mathbf{W}_{hx}\mathbf{X}_{t-1} + \mathbf{b}_h)$$

$$\mathbf{o}_t = \mathbf{W}_{ho}\mathbf{H}_t + \mathbf{b}_o$$

RNN with Gating: LSTMs and GRUs

RNN with Gating: LSTMs and GRUs

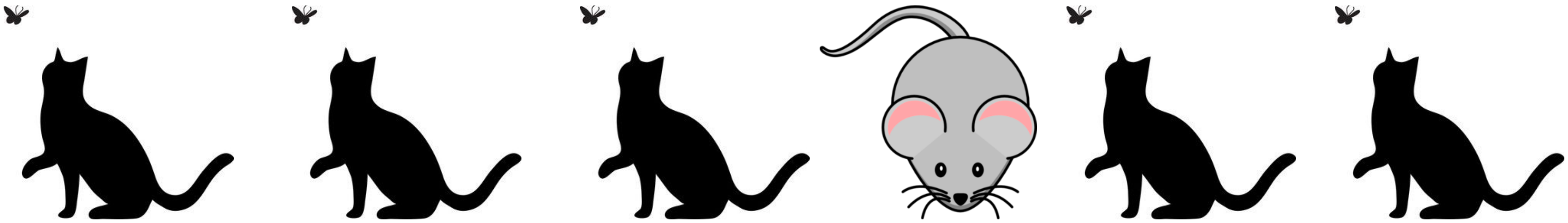
- Not all observations are equally relevant

RNN with Gating: LSTMs and GRUs

- Not all observations are equally relevant

RNN with Gating: LSTMs and GRUs

- Not all observations are equally relevant



- Only remember the relevant ones
  - Need mechanism to **pay attention (update gate)**
  - Need mechanism to **forget (reset gate)**

Generation with Neural Language Models

https://talktotransformer.com

Generation with Neural Language Models

GluonNLP is a deep learning framework for natural language processing and natural language processing. It is built on top of neural networks, and offers a robust, fast, and friendly solution for real-time text analysis and machine translation.

It comes packaged with some useful tools to help you with training and debugging, an editor so you can run your code in a safe and well-ventilated environment, and more.

https://talktotransformer.com

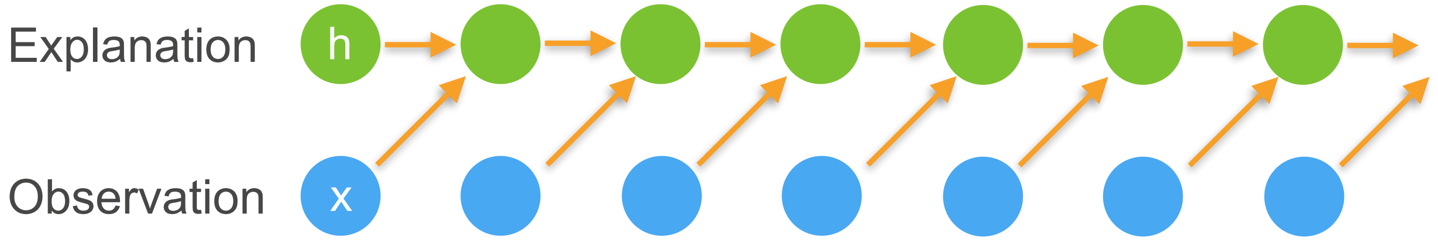
gluon-nlp.mxnet.io

Bi-Directional Models

- What is we have access to “future” context?

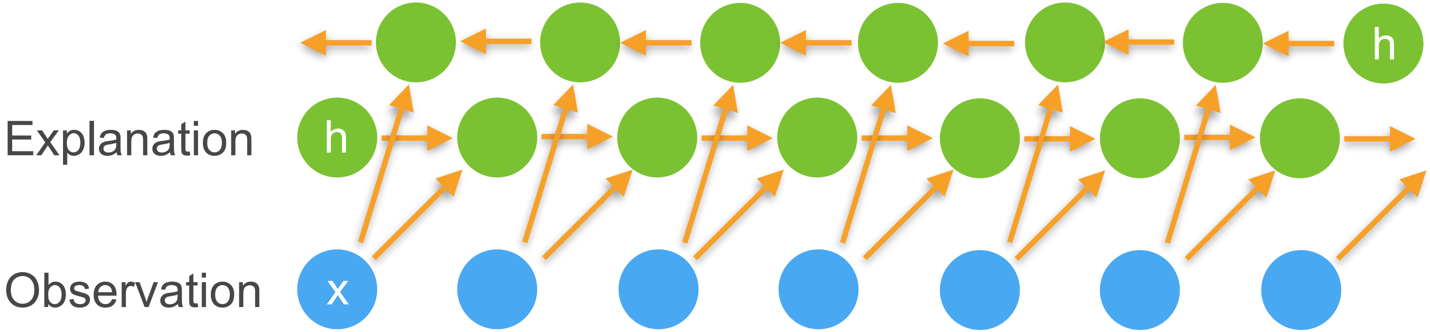
Bi-Directional Models

- What is we have access to “future” context?



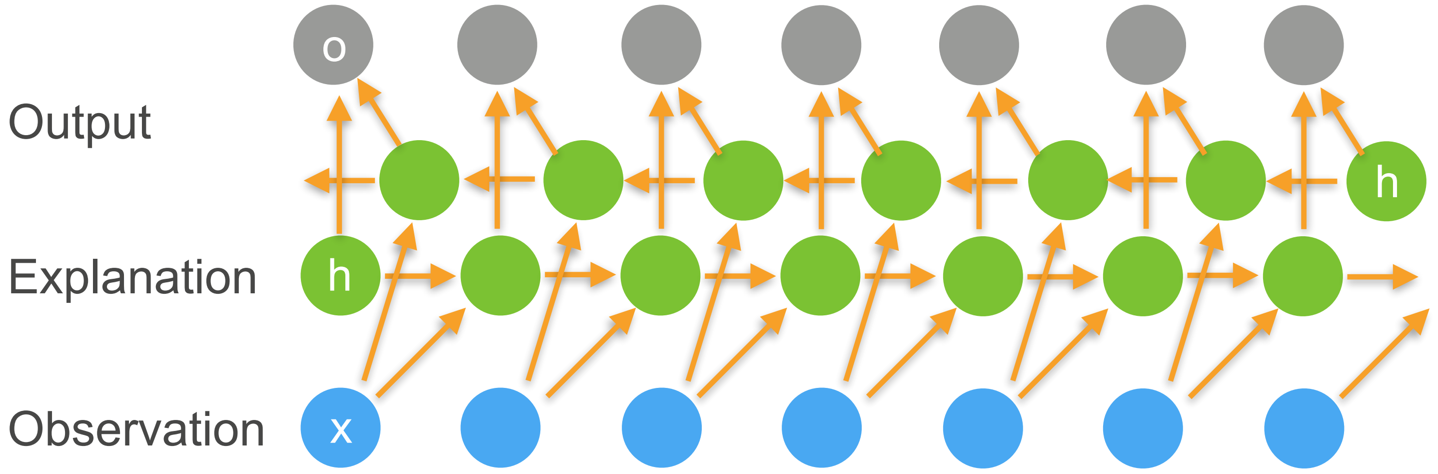
Bi-Directional Models

- What is we have access to “future” context?

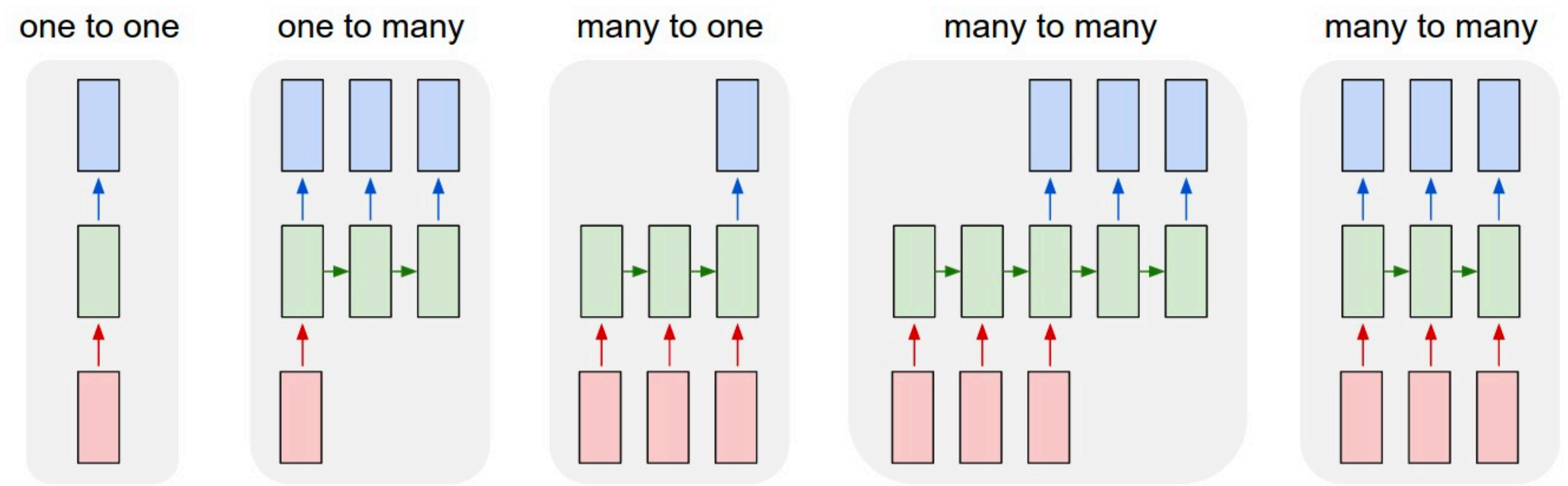


Bi-Directional Models

- What is we have access to “future” context?



Using RNNs



Poetry Generation

Sentiment Analysis

Document Classification

Question Answering

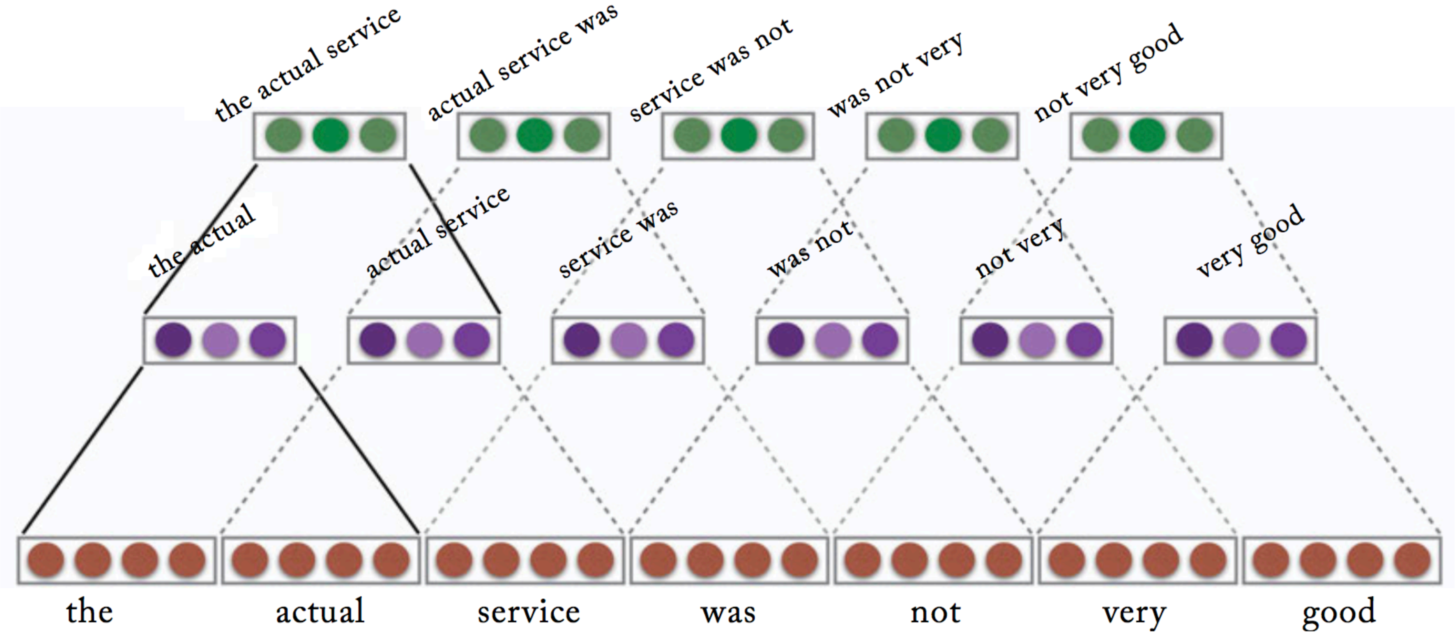
Machine Translation

Named Entity Tagging

(image courtesy of karpathy.github.io)

Convolutional Neural Networks

- Local to global



(image courtesy of Yoav Goldberg)

Subword Models

Subword Models

- What is a “word”?
  - 深度学习如何入门?
  - Lebensversicherungsgesellschaftsangestellter

Subword Models

- What is a “word”?
  - 深度学习如何入门?
  - Lebensversicherungsgesellschaftsangestellter
- Large vocabulary -> large model -> slow training

Subword Models

- What is a “word”?
    - 深度学习如何入门?
    - Lebensversicherungsgesellschaftsangestellter
- Large vocabulary -> large model -> slow training
- Out-of-Vocabulary (OOV) words
    - Loooooooooove
    - langauge

Tokenization

- Basic Idea - map text into sequence of tokens
  - “Deep learning is fun” -> [“Deep”, “learning”, “is”, “fun”, “.”]
- **Character Encoding** (each character as a token)
  - Small vocabulary
  - Doesn’t work so well (needs to learn spelling)
- **Word Encoding** (each word as a token)
  - Accurate spelling
  - Doesn’t work so well (huge vocabulary = costly multinomial)
- **Byte Pair Encoding** (Goldilocks zone)
  - Frequent subsequences (like syllables)