Natural Language Processing Basics

Colin Magdamo

Why is NLP Hard?

- “I struggle with the anxiety and it can feel incredibly crippling to the point I don’t want to face people because I feel I have commited a crime just by thinking one of my colleagues was laughing at me or because I said to my manager I think my colleague is trying to get rid of me… I wish I never acted on those thoughts because I am now struggling to go back to the workplace becase I feel ashamed…But I don’t want to completely disspear. I love life, I just want to embrace it.”
- Deceptively Complex
    - Several Agents: Author, colleague, manager. Who did what in this sentence?
- Typos and syntax errors, abbreviations
- Subtle inflection, slang, and implied meaning
    - Committed a crime isn’t literal here, but how would a computer know that?
    - “Acted” is used differently than “he acted in the play”
- Subjective Interpretation
    - What would you rate this person’s happiness on a scale of 1-5?

Modeling Using Language Derived Features

- Cox Proportional Hazards Model: $\lambda(\mathrm{t}|X_i) = \lambda_0(\mathrm{t}) \cdot e^{X_i \cdot \beta}$
  - Where X_i are the realized covariates for subject i
  - In our dementia survival model, we model the hazard with covariates like stroke before baseline as defined by the presence of ICD codes pertaining to stroke as defined by Mark.
  - Instead, we could use a cruder system that looks for exact match of "stroke" in any of the given subject's notes. This saves Mark time, but is obviously inaccurate:
    - "stroke" could be used in the following contexts (plus an infinite array of others):
      - "Patient's father had a history of stroke" – contained in a PCP note
      - "Patient had an ischemic stroke" – contained in ICU note
      - "Patient has difficulty with their breastroke" – contained in PT notes

Regex: Regular Expressions

- A mini programming language to specify patterns of text using rules
- *[\s]stroke
- (ischemic|hemorrhagic)[\s]stroke
- Better, but what about the following in a note:
  - “Patient has been playing more golf recently and even commented on taking a stroke off their game”
  - Clearly, we need to have some notion of *meaning* if we hope to usefully automate the process for extremely nuanced tasks

Language Models

The patient has dementia. They constantly seem to misplace their

History

WORD	PROB
keys	40%
wallet	35%
smartphone	5%
confidence	5%
jackhammer	.0000001%
…..	……

Prediction

Previous Ways of Representing Language as ML Input

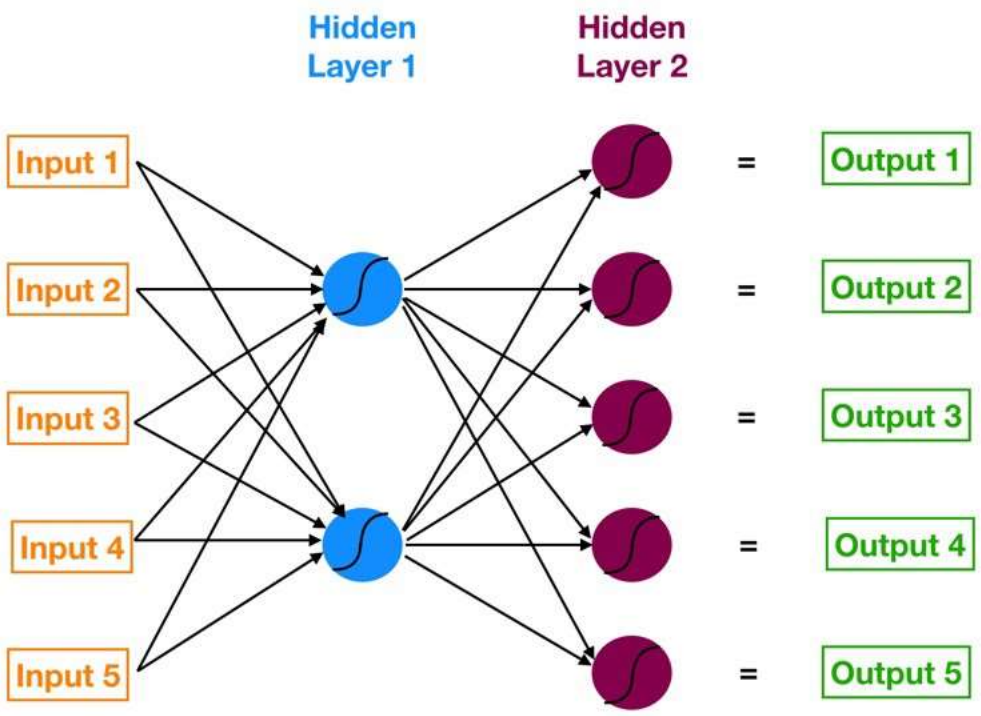
- Counts of Fixed Terms
  - One-Hot Encoding
- Counts of Regular Expressions
- Counts of N-Grams
- Word Vectors
  - Pretrained Fixed Word Vectors
    - Word2Vec
    - GloVe

Word Vectors: GLOVE

Problems With Representations

- One-Hot
  - Dot Product is always 0
- Regex
  - Hard to know your regex in advance, meaning comes from human design
- Counts of N-Grams
- Pretrained Word Vectors
  - Word will get the same vector regardless of context

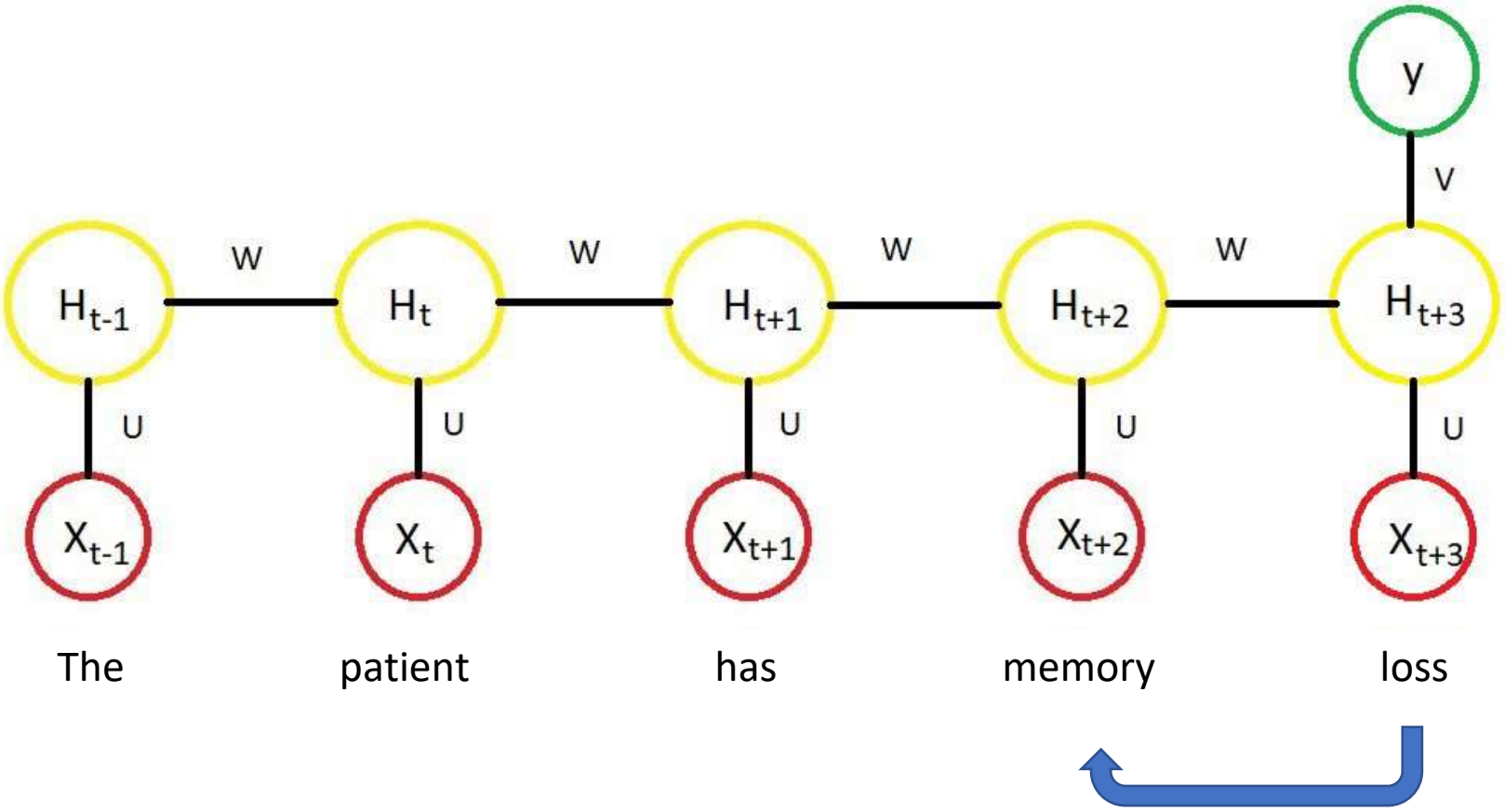
Vanilla Neural Networks



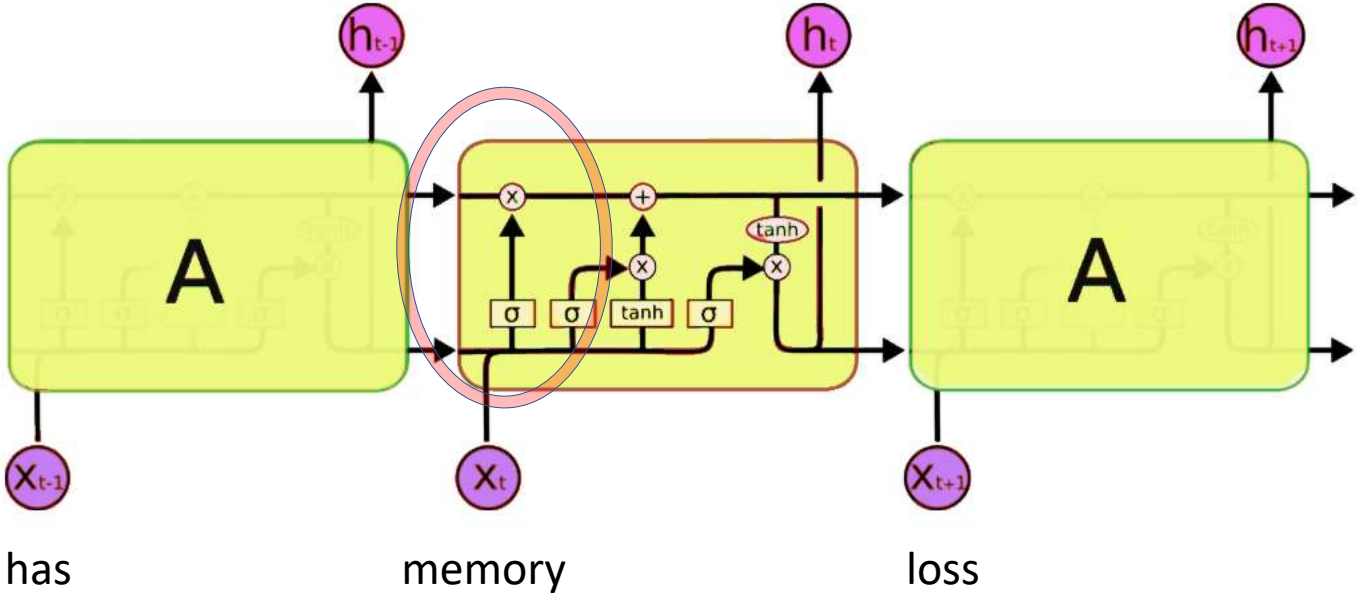
Why Do Vanilla Neural Networks Fail for Language Tasks?

- They don’t take into account any sequential information whatsoever; if we supply linguistic features, we get no deeper context surrounding the individual words or tokens in a piece of text.
- For something like classifying dementia, where there is both local sequence information (sentence level) and staging on a large time scale (units in years), this is a huge drawback
- “Memory of losing husband” means something very different from “husband losing memory”
  - Extra complication: losing here could mean losing track of in a grocery store or death of husband

(Old) Solution: Recurrent Neural Networks (RNNS)



(Newer, But Still Old Solution): Long Short Term Memory Networks (LSTM)

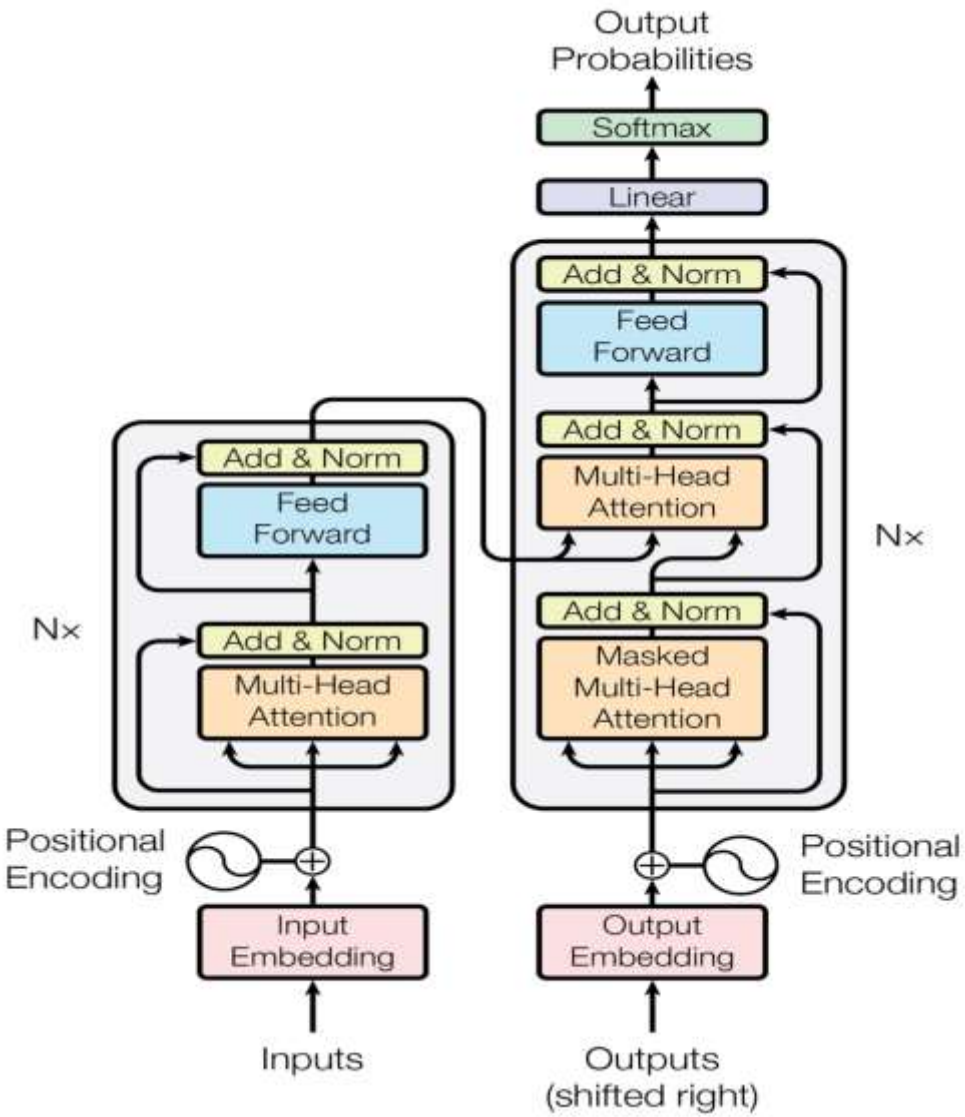


The extremely powerful ability to decide to forget! Imagine if this sequence of “has memory loss” was 100 words after “The patient”.

Problems with RNN Architectures

- Asymmetric: Can only read from one direction
  - Words near the start of sequence only get information about the start
- Only get a “scratchpad” memory of finite length
  - How do you know in advance what will and will not matter? The memory management problem balloons with size of text (patient note is much larger in scope than a tweet), and complexity of concept being described (language space in chart has many more ambiguous acronyms and references than a yelp review)
- “We decided to treat them with conservative management. Their father had a history of depression. They are taking an SSRI. They exhibited some signs of memory loss and confusion and seemed especially down.”
  - How does a computer know conservative is not being used in a political context?
  - How does it know down is the affect of the patient and not physical orientation?
  - How does it know the last clause is referring to the “them” (the patient) and not “their father”?
  - How does it know that history of depression is related to “seeming down”?
  - How does it know the SSRI is a medication and that taking means to ingest, not to steal?
  - How does it know the SSRI is related to being depressed?

Finally, the Modern Solution: Attention Mechanisms and the Transformer



https://arxiv.org/abs/1706.03762

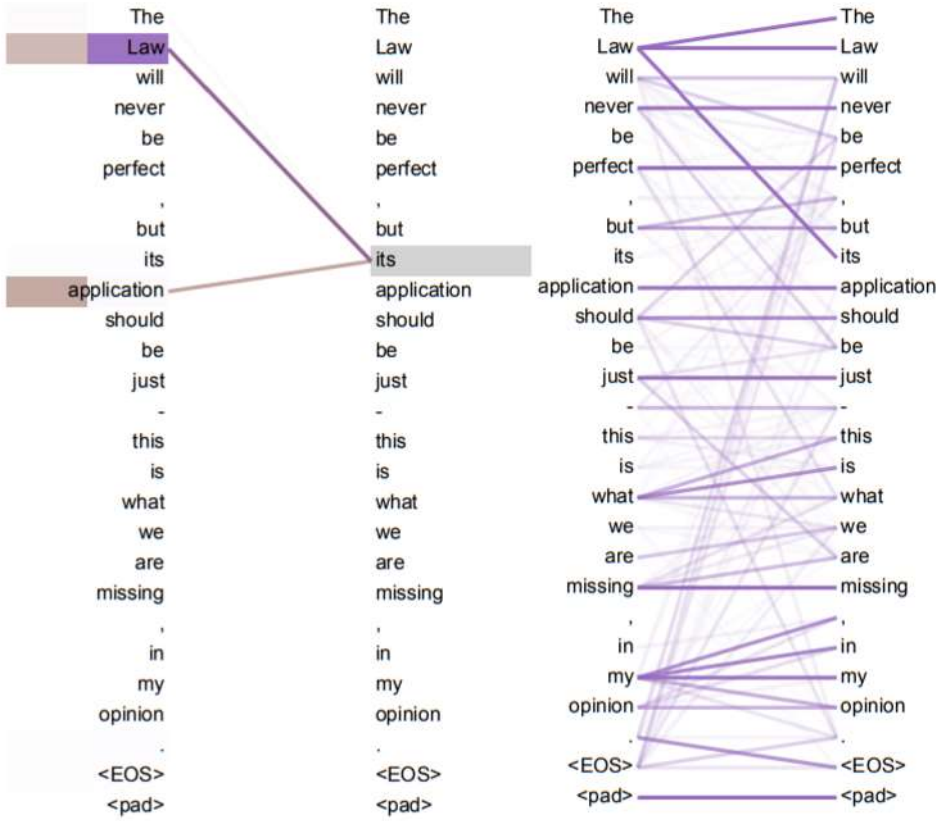
GPT2: Scarily Good Sequence Generation

- Prompt (History): The patient has dementia.
- Output: Approximately 45% of the population is predisposed to this disease by some unknown factor (such as high alcohol consumption). That being said, none of the treatments available will cure Alzheimer's disease, so instead, what I've learned is that putting people on medication for depression, anxiety, and even bipolar disorder with inadequate diet will simply make the condition worse. So the first step in treating this chronic condition is making a person's diet and lifestyle more conducive to health. The easiest way to do this is to stop drinking alcohol.
  If you're a regular drinker, the best thing you can do for yourself is stop.
  Another way to reverse the effects of drinking too much alcohol is to stop.

Attention: Overview

- Core idea: not all words contribute equally to sentence meaning
  - Assign a weight to each element of input sequence
  - Recurrence is mostly thrown out the window, and instead a positional encoding is learned
- Certain words derive their meaning from surrounding words, and these meanings interact to form complex higher level concepts such as this piece of text is describing a patient with dementia
- Encoder – Decoder Architecture
  - Encode sequence into a new vector, then decode that vector into another sequence

But What Does Attention Give Me?



How is it Used?

- Sequence to Sequence Translation
- Classification (our most relevant task)
- Dependency Parsing
- Sentiment Analysis
- Essentially, most modern NLP tasks have an attention mechanism somewhere underneath the hood
- Can be used in tandem with vanilla NN or RNN