Deep Learning (1470)

Randall Balestriero

Class 10: Sequential Data and Language Modeling

Recap!

Recap!

- What is dropout?

Recap!

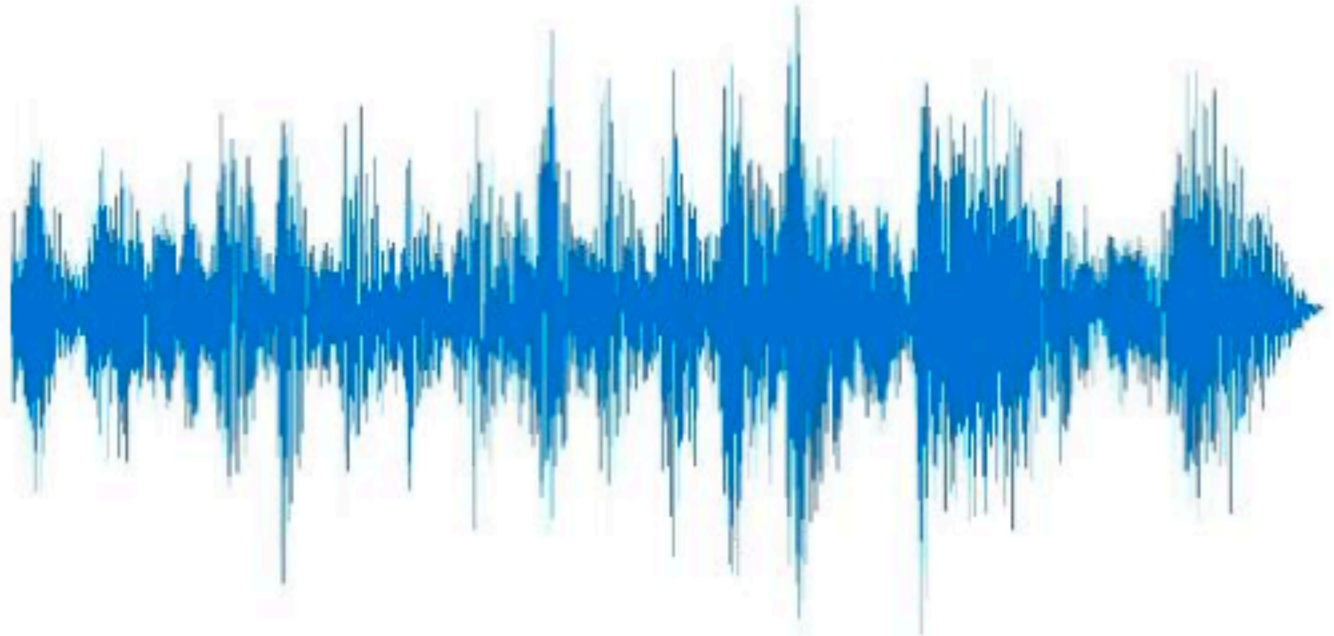
- What is dropout?
- What is drop path?

Recap!

- What is dropout?
- What is drop path?
- Why do we need to learn about residual connections and batch norm?

Sequential data

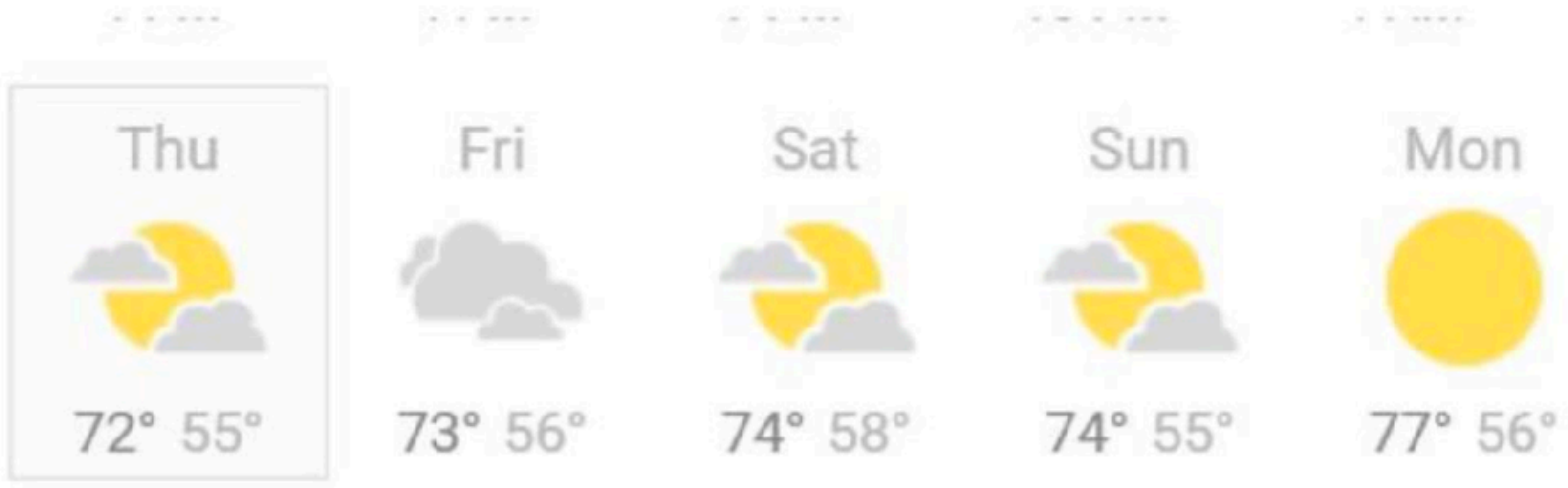
- Audio



- DNA

- Stock market

- Weather



Natural Language

Natural Language

- Sequence of words

Natural Language

- Sequence of words
- *"They went to the grocery store and bought bread, peanut butter, and jam."*

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, peanut butter, and jam."***
- Can be used for classification tasks

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, peanut butter, and jam."***
- Can be used for classification tasks
    - Sentiment analysis

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, peanut butter, and jam."***
- Can be used for classification tasks
  - Sentiment analysis
  - Spam detection

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, peanut butter, and jam."***
- Can be used for classification tasks
  - Sentiment analysis
  - Spam detection
- Can be used for generative tasks

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, peanut butter, and jam."***
- Can be used for classification tasks
    - Sentiment analysis
    - Spam detection
- Can be used for generative tasks
    - Content creation

Natural Language

- Sequence of words
- *"They went to the grocery store and bought bread, peanut butter, and jam."*
- Can be used for classification tasks
    - Sentiment analysis
    - Spam detection
- Can be used for generative tasks
    - Content creation
    - Assistant

Natural Language

- Sequence of words
- ***"They went to the grocery store and bought bread, pe nut butter, and jam."***
- Can be used for classificatio
    - Sentiment an
    - Spam detecti
- Can be used for generative tasks
    - Content creation
    - Assistant

Reality is more nuanced: translation?

Language Modeling

How to represent language: tokenization

“They went to the grocery store and bought bread, peanut butter, and jam.”

- Consistent casing
- Strip punctuation
- One word is one token
- Split on spaces

```
[“they”, “went”, “to”, “the”,
“grocery”, “store”, “and”,
“bought”, “bread”, “peanut”,
“butter”, “and”, “jam”]
```

Language Modeling

How to represent language: tokenization

- Choose a hyperparameter vocab_size for how many words the model should know

- Keep only the vocab_size most frequent words and replace everything else with [UNK]

Language Modeling

How to represent language: tokenization

- Choose a hyperparameter vocab_size for how many words the model should know

- Keep only the vocab_size most frequent words and replace everything else with [UNK]

– *"They galloped to the Ratty for dinner, and ate exactly seventy-three waffle fries and chocolate peamilk."*

Language Modeling

How to represent language: tokenization

- Choose a hyperparameter vocab_size for how many words the model should know

- Keep only the vocab_size most frequent words and replace everything else with [UNK]

- *"They galloped to the Ratty for dinner, and ate exactly seventy-three waffle fries and chocolate peamilk."*

- ["they", "UNK", "to", "the", "UNK", "for", "dinner", "and", "ate", "exactly", "UNK", "waffle", "fries", "and", "chocolate", "UNK"]

Language Modeling

How to represent language: tokenization

- Choose a hyperparameter vocab_size for how many words the model should know
- Keep only the vocab_size most frequent words and replace everything else with [UNK]
- More complicated tokenization strategies: can you think of another example?

Language Modeling

How to model language: conditional probability

- $p(\text{token}_1, \text{token}_2, \text{token}_3) = p(\text{token}_1)p(\text{token}_2 \mid \text{token}_1)p(\text{token}_3 \mid \text{token}_1, \text{token}_2)$

Language Modeling

How to model language: conditional probability

- $p(\text{token}_1, \text{token}_2, \text{token}_3) = p(\text{token}_1)p(\text{token}_2 \mid \text{token}_1)p(\text{token}_3 \mid \text{token}_1, \text{token}_2)$

P("*they went to the store*") = P("*they*")*P("*went*"|"*they*")*P("*to*"|"*they went* ")* ..

Language Modeling

How to model language: conditional probability

- $p(\text{token}_1, \text{token}_2, \text{token}_3) = p(\text{token}_1)p(\text{token}_2 \mid \text{token}_1)p(\text{token}_3 \mid \text{token}_1, \text{token}_2)$

P(“*they went to the store*”) = P(“*they*”)*P(“*went*”|“*they*”)*P(“*to*”|“*they went* ”)* ..

What is the size of the transition matrix?

Language Modeling

How to model language: conditional probability

- $p(\text{token}_1, \text{token}_2, \text{token}_3) = p(\text{token}_1)p(\text{token}_2 \,|\, \text{token}_1)p(\text{token}_3 \,|\, \text{token}_1, \text{token}_2)$

P("*they went to the store*") = P("*they*")*P("*went*"|"*they*")*P("*to*"|"*they went* ")* ..

What is the size of the transition matrix?

Quickly becomes intractable and with most sequences having 0 probability

Language Modeling

How to model language: conditional probability

- Goal: predict next word given a preceding sequence
  - $P(\boldsymbol{word_n} | \, word_1, word_2, \ldots word_{n-1}) = \frac{Count(word_1, word_2, \ldots word_{n-1}, \boldsymbol{word_n})}{Count(word_1, word_2, \ldots word_{n-1})}$
- Example task: predict the next word
  - ***he danced*** ___
- Strategy: iterate through all words in vocabulary, and calculate

  $\frac{Count(he\ danced\ <word>)}{Count(he\ danced)}$ for each word

Language Modeling

How to model language: conditional probability

- Our training sentences were:
    - *"She danced happily"*
    - *"They sang beautifully"*
    - *"He danced energetically"*
    - *"He sang happily"*
    - *"She danced gracefully"*

$$\frac{Count(he\ danced\ < word >)}{Count(he\ danced)}$$

- *"He danced ___ "*

- *"He danced* ***happily***"

Has 0 probability

Language Modeling

How to model language: conditional probability

Improvement: **N-gram** model – only look at **N** words at a time

(in this case, **bi**grams look at **2** words at a time)

–*"danced happily"*
–*"sang beautifully"*
–*"danced energetically"*
–*"sang happily"*
–*"danced gracefully"*

"He danced happily" now has 1/3 probability!

But what if the answer was *"He danced beautifully"* ?

Language Modeling

How to model language: conditional probability

Improvement: **N-gram** model – only look at **N** words at a time

(in this case, **bi**grams look at **2** words at a time)

-*"danced happily"*
-*"sang beautifully"*
-*"danced energetically"*
-*"sang happily"*
-*"danced gracefully"*

"He danced happily" now has 1/3 probability!

But what if the answer was *"He danced beautifully"* ?

Language Modeling

How to model language: conditional probability

Improvement: **N-gram** model – only look at **N** words at a time

(in this case, **bi**grams look at **2** words at

Let's use a Deep Network!

–“*danced ha*

–“*sang bea*

–“*danced e*

–“*sang happ*

–“*danced gracefully*”

“*He danced <u>happily</u>*” now has 1/3 probability!

But what if the answer was “*He danced <u>beautifully</u>*” ?

Language Modeling

How to model language: Deep Networks

Language Modeling

How to model language: Deep Networks

- We can model $p(\text{token}_t \mid \text{token}_1, \ldots, \text{token}_{t-1}) = f_\theta(\text{token}_1, \ldots, \text{token}_{t-1})$

Language Modeling

How to model language: Deep Networks

- We can model $p(\text{token}_t \mid \text{token}_1, \ldots, \text{token}_{t-1}) = f_\theta(\text{token}_1, \ldots, \text{token}_{t-1})$
- Is that a regression or a classification task?

Language Modeling

How to model language: Deep Networks

- We can model $p(\text{token}_t \mid \text{token}_1, \ldots, \text{token}_{t-1}) = f_\theta(\text{token}_1, \ldots, \text{token}_{t-1})$
- Is that a regression or a classification task?
- How many classes do we have?

Language Modeling

How to model language: Deep Networks

- We can model $p(\text{token}_t \,|\, \text{token}_1, \ldots, \text{token}_{t-1}) = f_\theta(\text{token}_1, \ldots, \text{token}_{t-1})$
- Is that a regression or a classification task?
- How many classes do we have?
- What do you think is a good architecture?

Questions?