IS883: Synthesizing Digital Efforts

Mohannad Elhamod

BOSTON UNIVERSITY

Language Modeling

BOSTON UNIVERSITY

Boston University Questrom School of Business

What does a model mean?

Britannica

Search Britannica...

scientific modeling

Table of Contents

Introduction

Fast Facts

Related Content

Media

Images

More

More Articles On This Topic

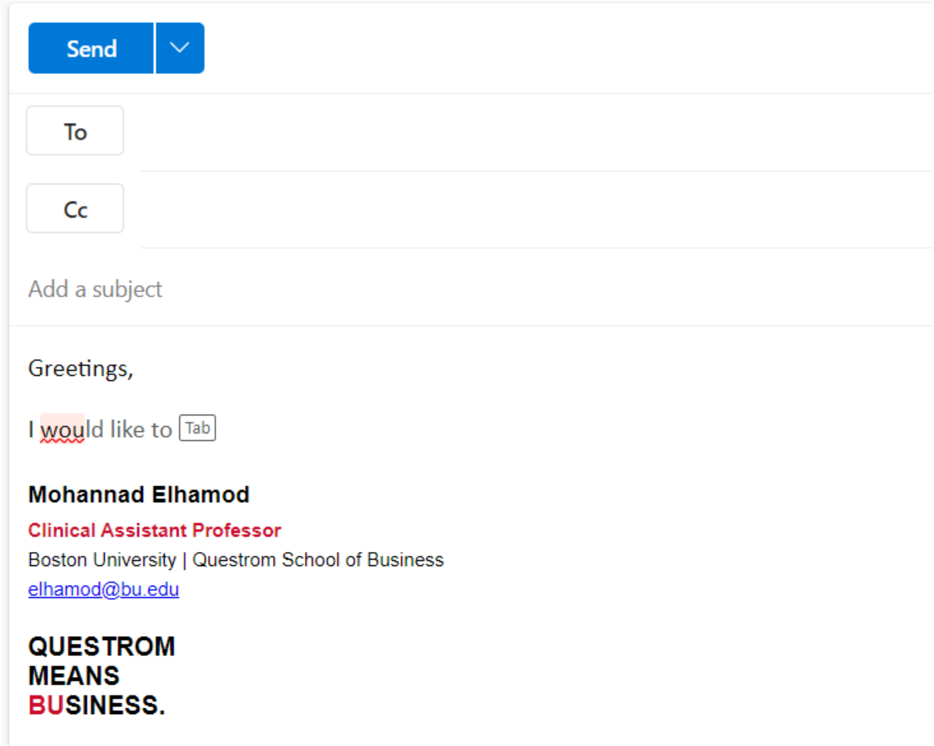
Contributors

Article History

scientific modeling, the generation of a physical, conceptual, or mathematical representation of a real phenomenon that is difficult to observe directly. Scientific models are used to explain and predict the behaviour of real objects or systems and are used in a variety of scientific disciplines, ranging from physics and chemistry to ecology and the Earth sciences. Although modeling is a central component of modern science, scientific models at best are approximations of the objects and systems that they represent—they are not exact replicas. Thus, scientists constantly are working to improve and refine models.

What is language modeling anyway?...

Web search engine / ...



Lena-voita

What is language modeling anyway?...

I grabbed the branch and broke it.

I went to the branch and deposited some money.

Context matters!

What is language modeling anyway?...

- I went to ____.
- I woke up at 7 am and went to ___.
- I woke up at 7 am, packed my book and notebook, and went to ____.

The more context, the more certain

What is language modeling anyway?...

I went to the branch and deposited some money.

I went to the bank and deposited some money.

I went to the ATM and deposited some money.

> Words which frequently appear in **similar contexts** have **similar meaning**.

Lena-voita

What is language modeling anyway?...

I sat at the bank and …

… watched the water flow.

… waited for my turn.

We process language sequentially*.

*We will talk about exceptions later...

Natural Language Processing (NLP)

Includes text generation:

- Text completion.
- Text summarization.
- Question answering.

But there are also many other tasks such as Text classification: (e.g., Sentiment analysis, Reviews, Fake news) or word classification.

Formalizing our thoughts

- So, language modeling is the chaining of word probabilities.

$$P(\text{I} \quad \text{saw} \quad \text{a} \quad \text{cat} \quad \text{on} \quad \ldots) =$$

$$\underbrace{P(\text{I}) \cdot P(\text{saw}|\text{I}) \cdot P(\text{a}|\text{I saw}) \cdot P(\text{cat}|\text{I saw a}) \cdot P(\text{on}|\text{I saw a cat})}_{\text{Probability of I saw a cat on}} \cdot \underbrace{\ldots}$$

Lena-voita

- How do we calculate these probabilities?

counting…

$$P(\text{cat}) = \frac{N(\text{"cat" in corpus})}{N(\text{all words in corpus})}$$

$$P(\text{cat} \mid \text{my}) = \frac{N(\text{"my cat" in corpus})}{N(\text{"my" in corpus})}$$

Can you foresee any problem with this calculation?...

N-grams

Instead, let's just use a context of ***fixed-length.***

P(I saw a cat on a mat) =

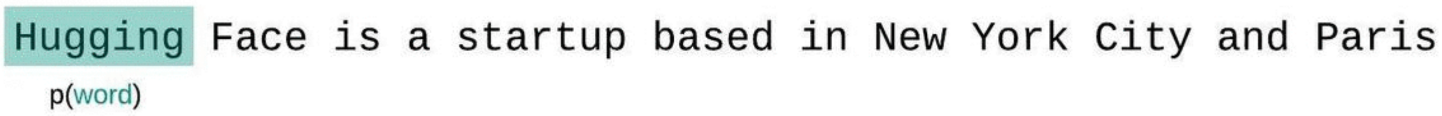
P(I)
· P(saw | I)
· P(a | I saw)
· P(cat | I saw a)
· P(on | I saw a cat)
· P(a | I saw a cat on)
· P(mat | I saw a cat on a)

Lena-voita

- n=3 (trigram model): $P(y_t|y_1, \ldots, y_{t-1}) = P(y_t|y_{t-2}, y_{t-1})$,
- n=2 (bigram model): $P(y_t|y_1, \ldots, y_{t-1}) = P(y_t|y_{t-1})$,
- n=1 (unigram model): $P(y_t|y_1, \ldots, y_{t-1}) = P(y_t)$.

N-grams

Context is like a sliding window into the past.



[Huggingface]

Context size

- I went to the beach. My wife sat next to me. She was replying to some emails, and the bird stole our sandwich. Then it started raining suddenly and ___.

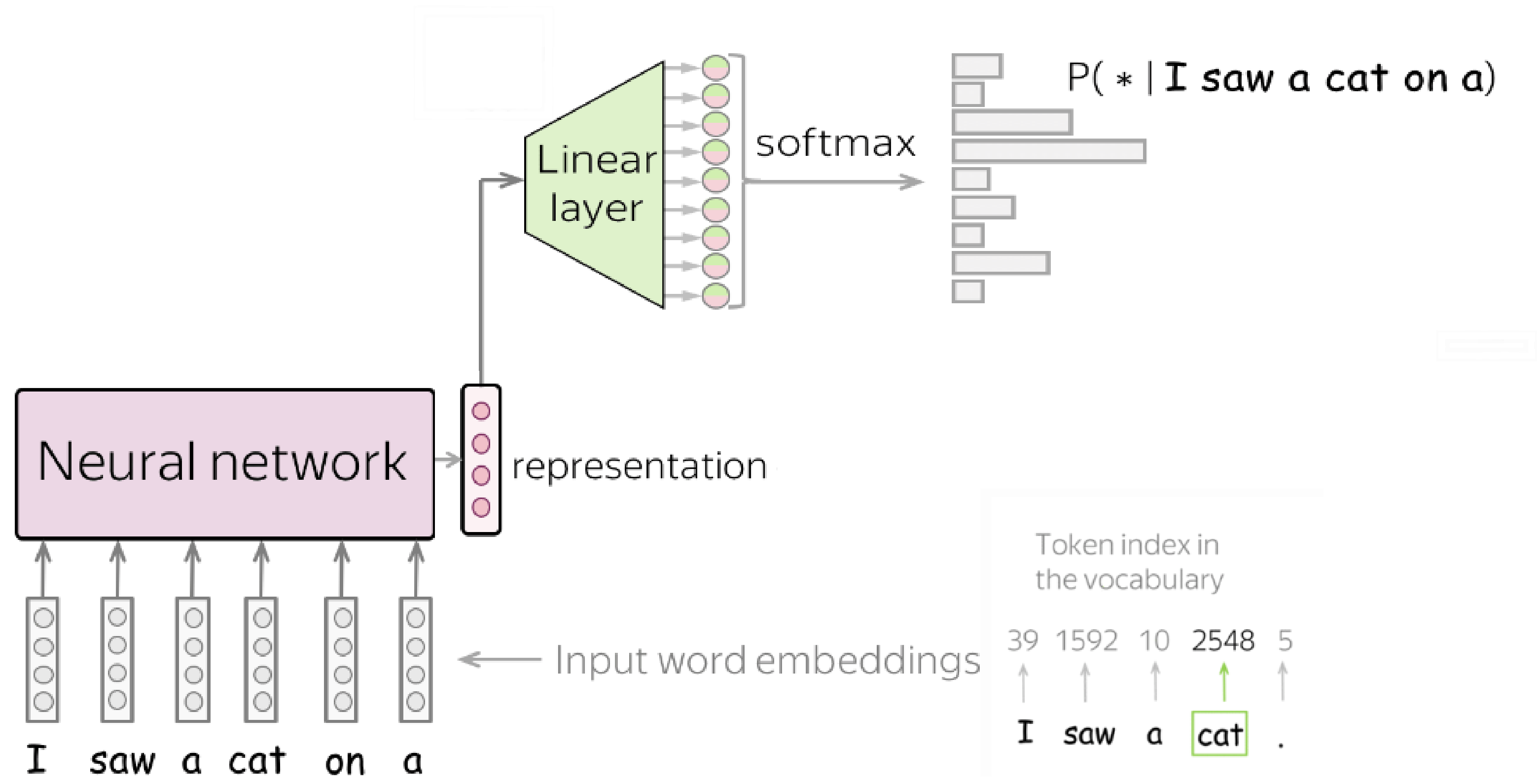
- Longer context: predictable outcome.
- Shorter context: Too unpredictable.

Neural networks for language modeling

Boston University Questrom School of Business

BOSTON UNIVERSITY

General Model Architecture



Lena-voita

Can you see any issue with inputting words in an NN?

What is an embedding?

- embeddings = representation= features = latent space.
- It is a representation of your input.
- [Example for images.]

Word embeddings

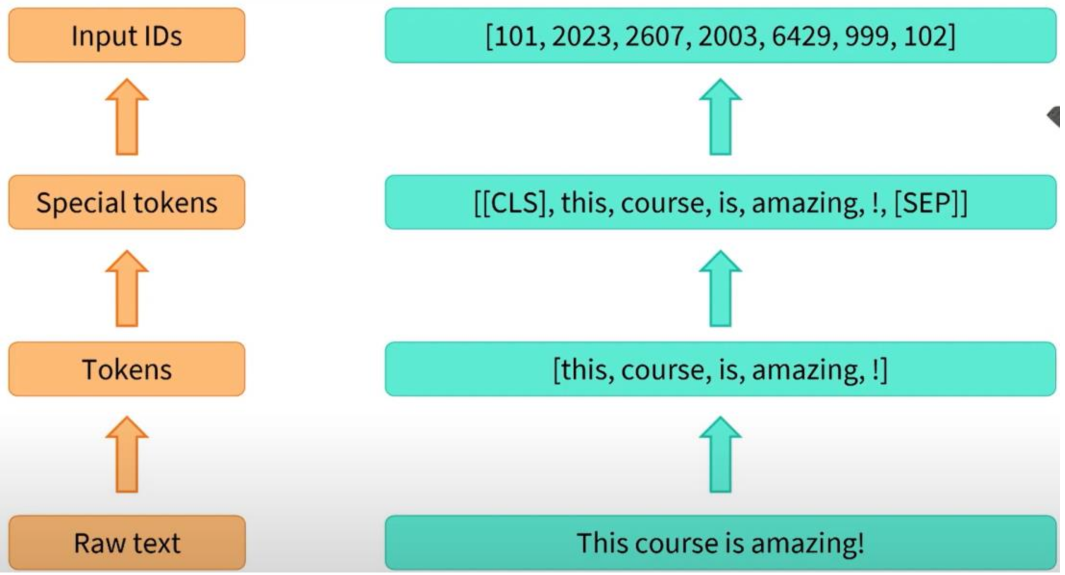
- We ideally want words that have similar meanings to have smaller distances.
- Demo
- Examples:
  1. Word2Vec (Google)
  2. GloVe (Stanford)
  3. Train your own!

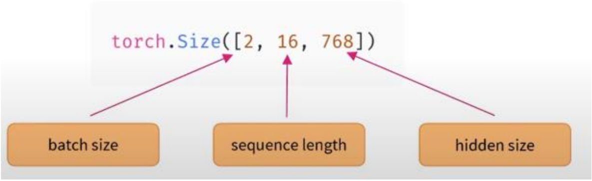
Tokenization

So, every time we have sentences to generate, we represent them as:

- A batch of sentences (i.e., batches)
- Each sentence is represented as a sequence of tokens (sequence length)
- Each token is represented as a vector (hidden size)

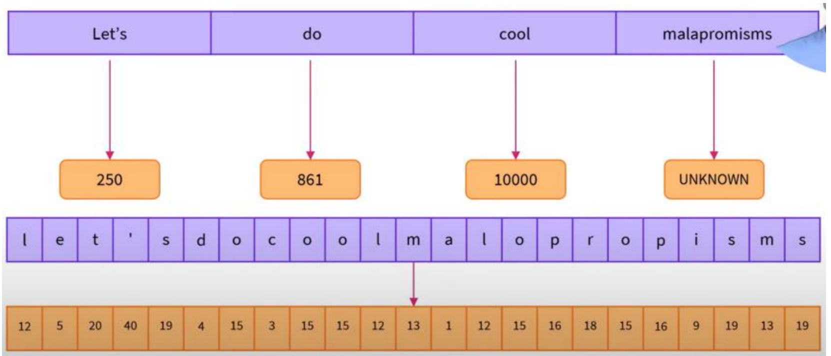
Huggingface

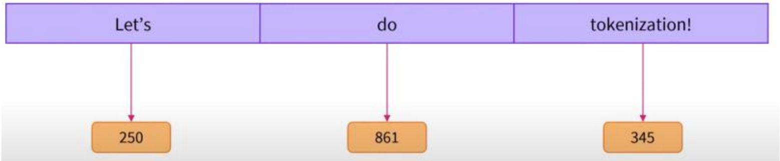




Why do word level?

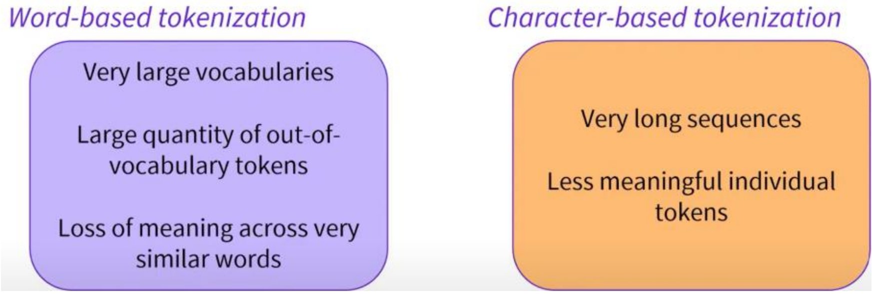
Huggingface



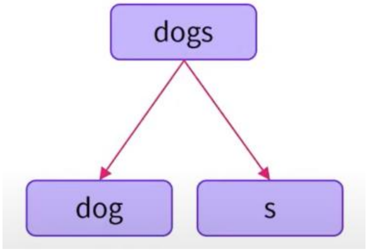


Issues?

Issues?

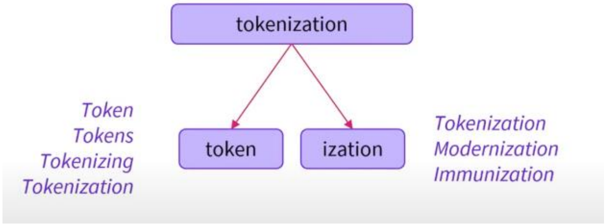


Why do word level?



How about sub-words?

- Preserves word morphology.
- Can represent new words.
- Handles misspelling.
- Examples:



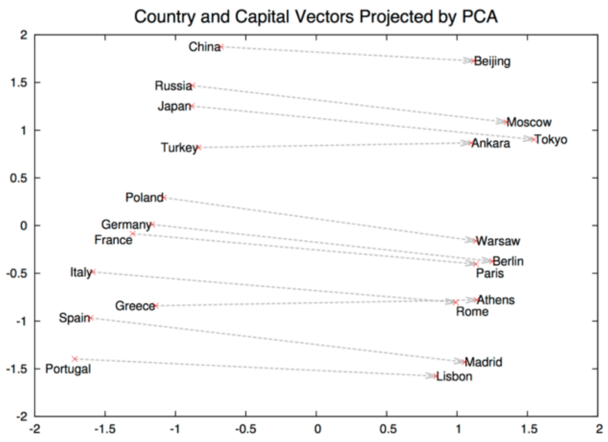
Word embeddings

Word embeddings can also be used find directionality in the corpus:

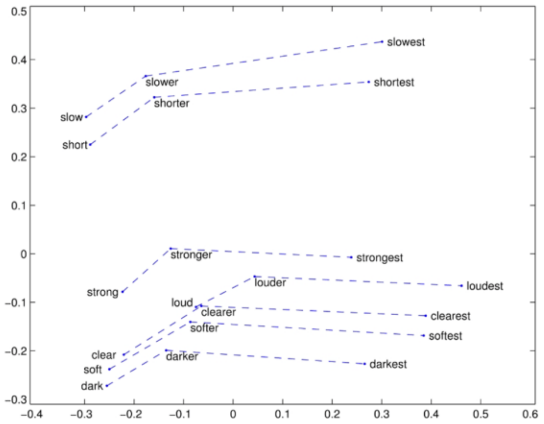
- Demo 1 (semantics)
- Demo 2 (vector view)
- Demo 3 (dimensionality)

Word embeddings

Word embeddings can also be used find directionality in the corpus.



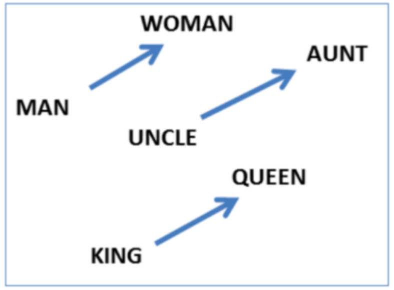
Word2Vec

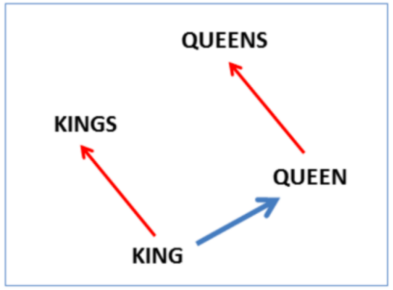


GloVe

semantic: v(king) – v(man) + v(woman) ≈ v(queen)

syntactic: v(kings) – v(king) + v(queen) ≈ v(queens)

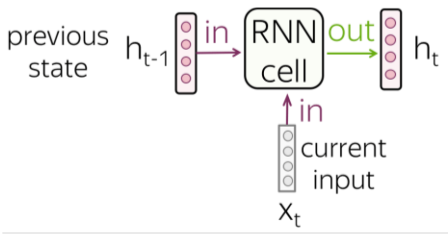


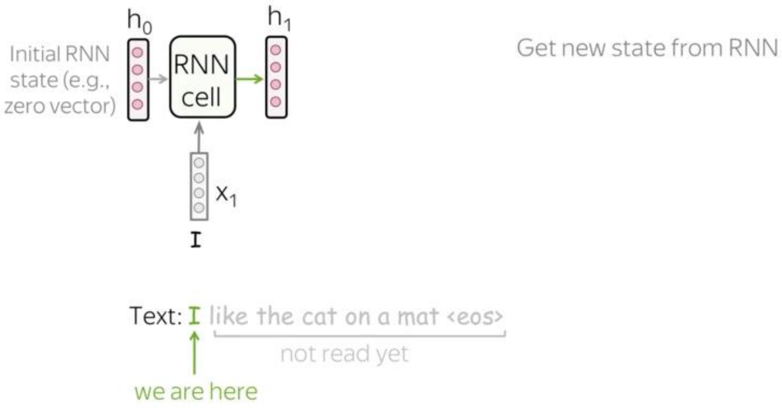


Recurrent Neural Nets (RNNs)

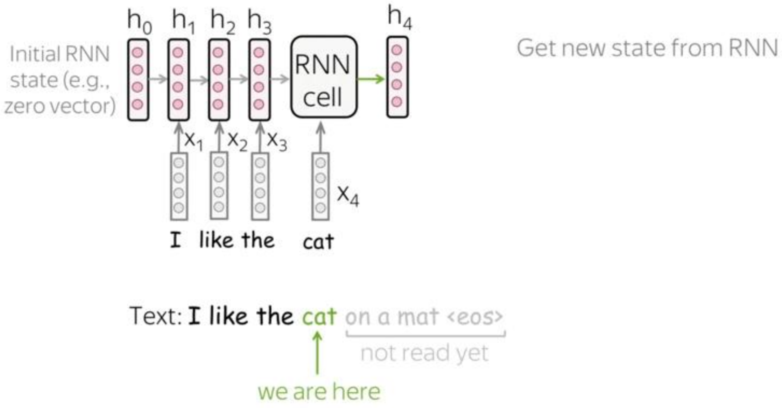
- Combines the embeddings of previous context and current word → gives next word.

Lena-voita



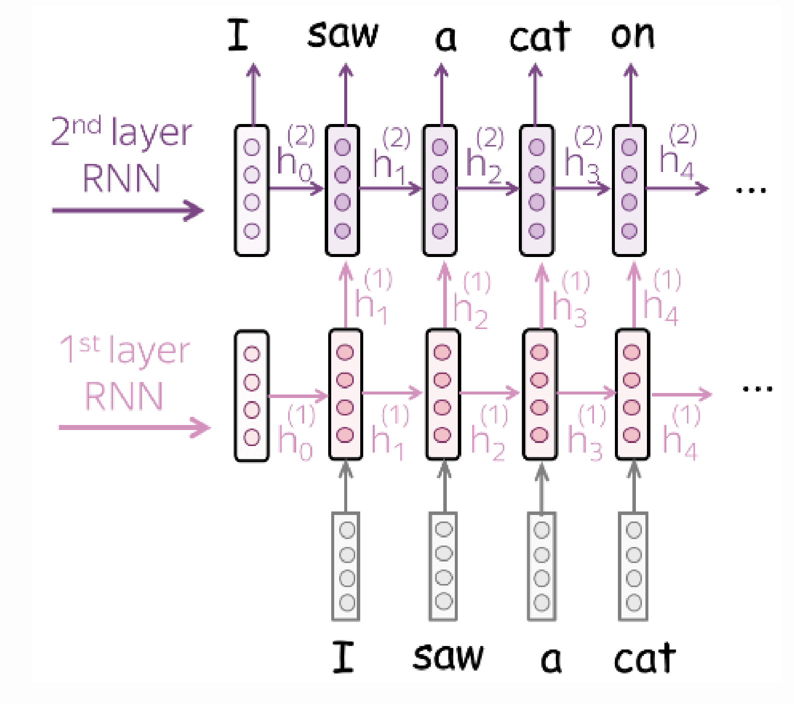


....



Recurrent Neural Nets (RNNs)

- We can add more layers and units per layer to increase complexity.



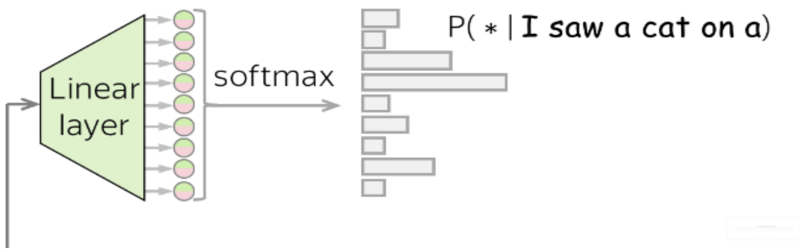
Lena-voita

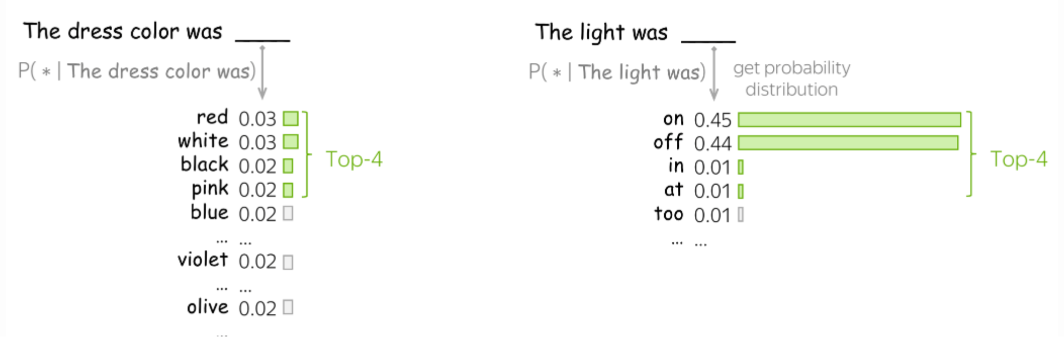
Recurrent Neural Nets (RNNs)

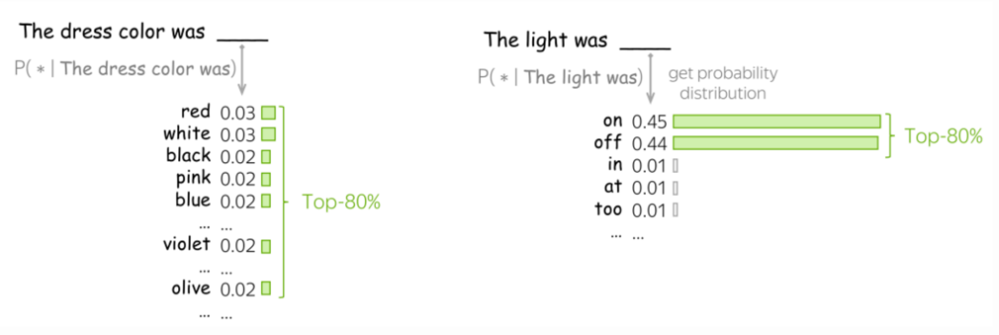
Demo

Sampling The Distribution

- Always take top probability?
  - That makes the model deterministic (no creativity).
- Alternative?
  - Top-k or top-p.



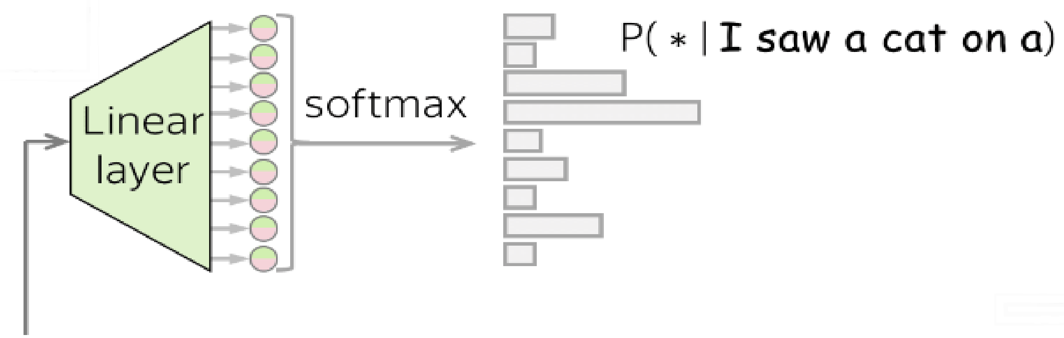


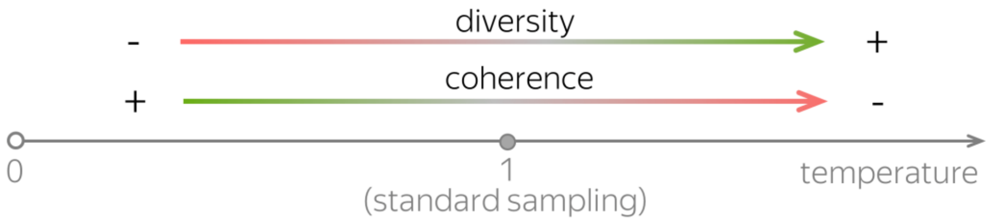


Lena-voita

Sampling The Distribution

- Some words have way higher probability than others.
- This can be manually tuned through temperature.
- Demo





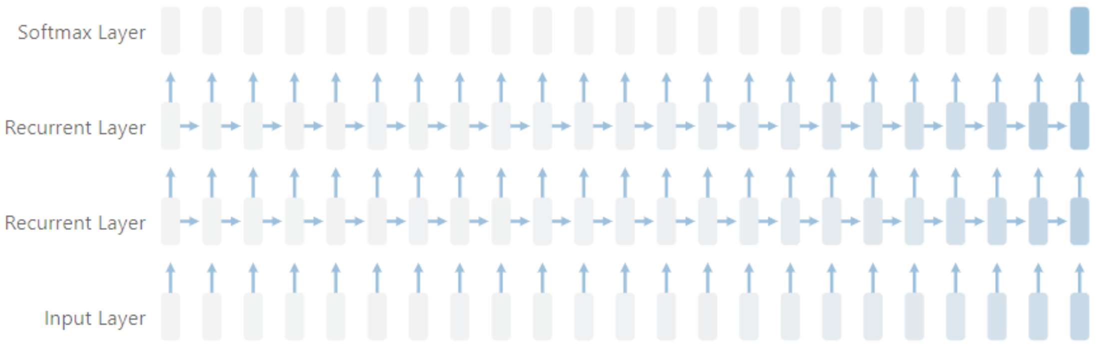
Lena-voita

Measuring The Metric

- What are we looking for?
  - A model that is not surprised by the new text it seen.
- We use perplexity.
  - Takes values between 1 and number of possible tokens.
  - Smaller is better.
  - Demo

RNNs (issues)

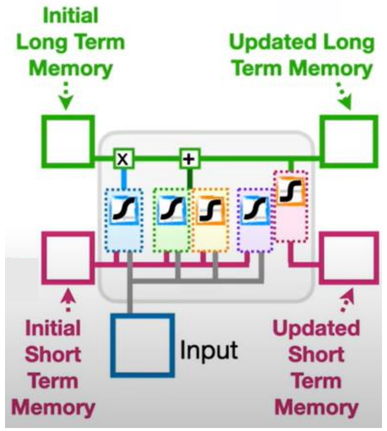
- Gradient becomes insignificant for long contexts
  - The network forgets early words…
  - It is called the “vanishing gradient” or “memorization” problem.
  - RNNs have an issue memorizing long contexts.



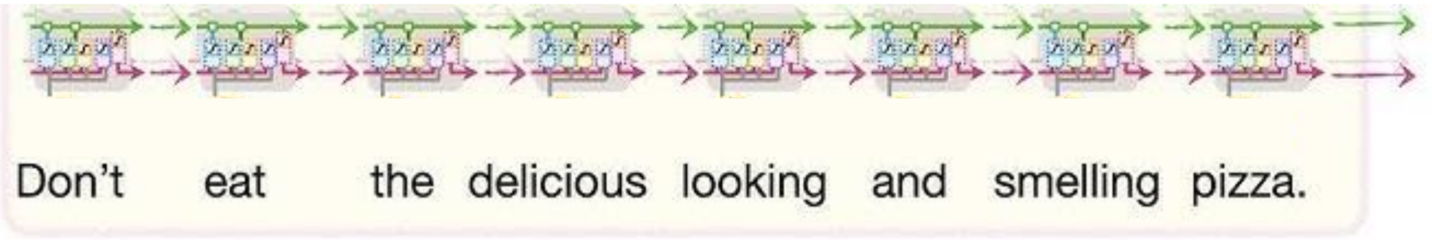
Vanishing Gradient: where the contribution from the earlier steps becomes insignificant in the gradient for the vanilla RNN unit.

distill.pub

Attempted solution: LSTM



- Instead of one representation, let's have two!
  - One for short-term memory
  - And one for long-term memory.
- Somewhat of an improvement.

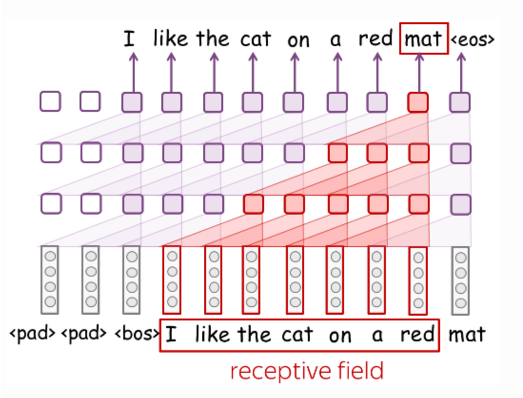


StatQuest

There are other similar models

GRUs, CNNs…

But we will not talk about them here.



CNN

Question for next week…

What limitations are inherent in traditional NLP models, and how might they be addressed?

Assignment