VerneBOT: GENERATING TEXTS LIKE *JULES VERNE*

An Introduction to Language Models

Prof. Marcelo Rovai, UNIFEI

VerneBOT

What is Verne Bot?

A model trained to generate text in the style of Jules Verne. He uses texts extracted from books such as *A Jorney to the Centre of the Earth* and *From the Earth to the Moon*.

Simplified introduction to Large Language Models (LLMs) such as GPT.

Generating Text with RNNs: The Jules Verne Bot - Notebook

DATA PREPARATION

Data was collected from 10 books by Jules Verne (5.8 million characters).

Preprocessing: Removal of irrelevant characters and structuring of the text for analysis.

Importance of clean data for training.

Project Gutenberg

'A Journey to the Centre of the Earth'
'An Antarctic Mystery'
'Around the World in Eighty Days'
'Five Weeks in a Balloon'
'From the Earth to the Moon'
'In Search of the Castaways'
'In the year 2889'
'Michael Strogoff'
'The Mysterious Island'
'Twenty Thousand Leagues under the Sea'

TOKENIZATION AND VOCABULARY

Conversion of text into numeric tokens.

Character-level tokenization: 123 unique characters.

Example: “The Project” ⟶ [122 52 69 66 1 48 79 76 71 66 64 81].

TOKENIZATION

```
['\n', ' ', '!', '"', '$', '&', "'", '(', ')', '*', '+', ',', '-',
 '.', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ':', ';',
 '=', '?', 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K',
 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X',
 'Y', 'Z', '[', ']', '_', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h',
 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u',
 'v', 'w', 'x', 'y', 'z', '§', '«', '°', '»', 'À', 'Á', 'Ã', 'Ç',
 'É', 'Ê', 'Í', 'Ó', 'Ú', 'à', 'á', 'â', 'ã', 'æ', 'ç', 'è', 'é',
 'ê', 'í', 'î', 'ñ', 'ò', 'ó', 'ô', 'õ', 'ú', 'û', 'œ', '–', '’'],
```

```
[31, 42, 40,  1, 30, 28, 46, 40, 48, 45, 45, 42,  0,  0, 36,  0,  0,
 31, 71,  1, 76, 65, 76, 77, 68, 71, 13,  0,  0, 48, 69, 57,  1, 70,
 71, 65, 76, 61,  1, 60, 61, 75, 76, 57, 75, 11,  1, 78, 65, 70, 60,
 71,  1, 60, 57,  1, 59, 65, 60, 57, 60, 61,  1, 72, 57, 74, 57,  1,
 71,  1, 32, 70, 63, 61, 70, 64, 71,  1, 41, 71, 78, 71, 11,  1, 61,
 70, 59, 71, 70, 76, 74, 61, 65,  1, 70, 71,  0, 76, 74, 61, 69,  1,
 60, 57,  1, 30, 61, 70, 76, 74, 57, 68,  1, 77, 69,  1, 74, 57, 72,
 57, 82,  1, 57, 73, 77, 65,  1, 60, 71,  1, 58, 57, 65, 74, 74, 71,
 11,  1, 73, 77, 61,  1, 61, 77,  1, 59, 71, 70, 64, 61])
```

https://platform.openai.com/tokenizer

TRAINING SEQUENCES

Goal: Predict the next character in a sequence.

Length of the sequence: 150 characters (paragraph).

Input 'Hello my nam'

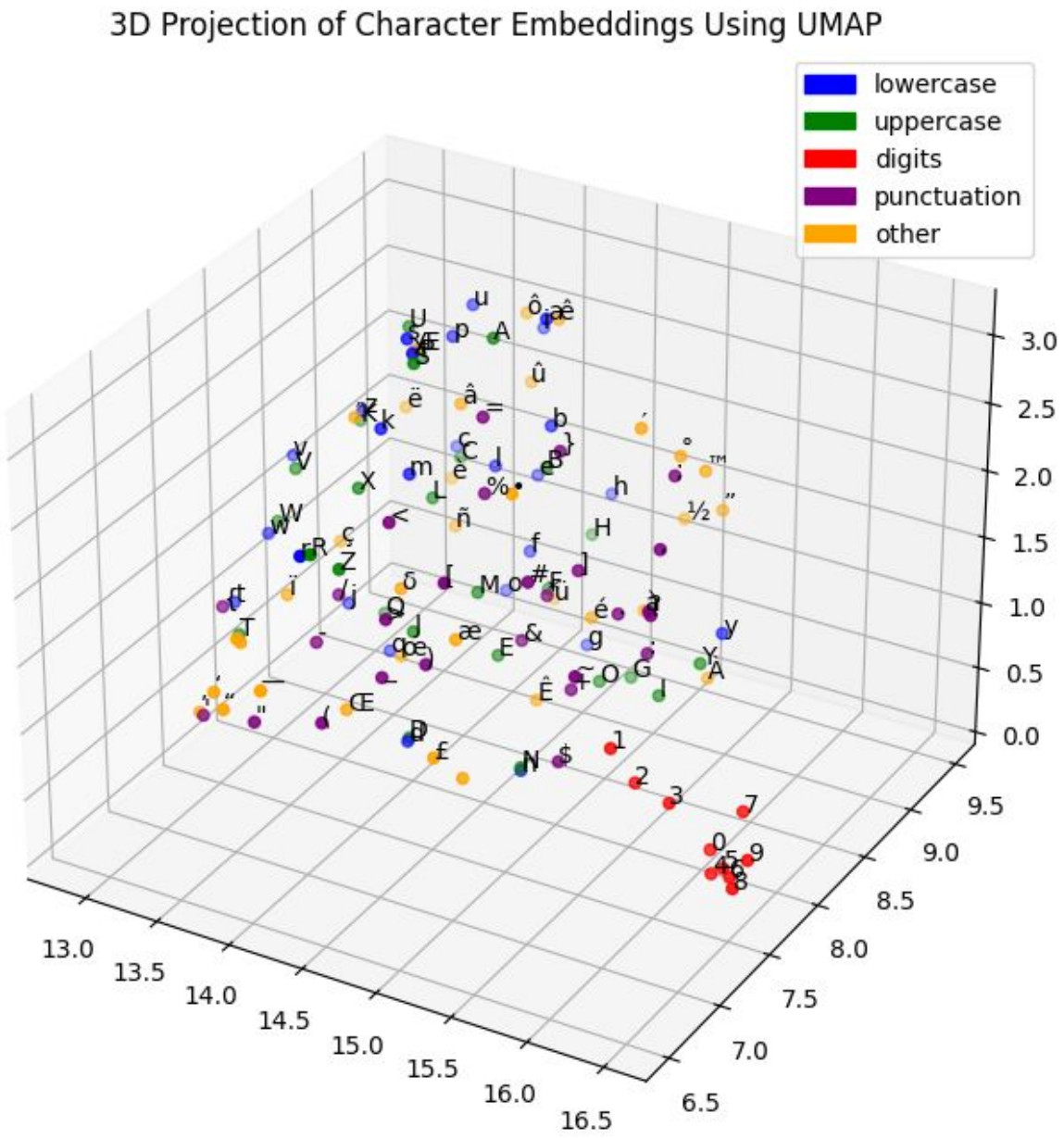
Output 'ello my name'.

EMBEDDING

Each character is represented as a vector of 256 dimensions.

Embedding captures relationships between characters in dense vectors.

EMBEDDING



Word2Vec - Embedding Projector

MODEL ARCHITECTURE

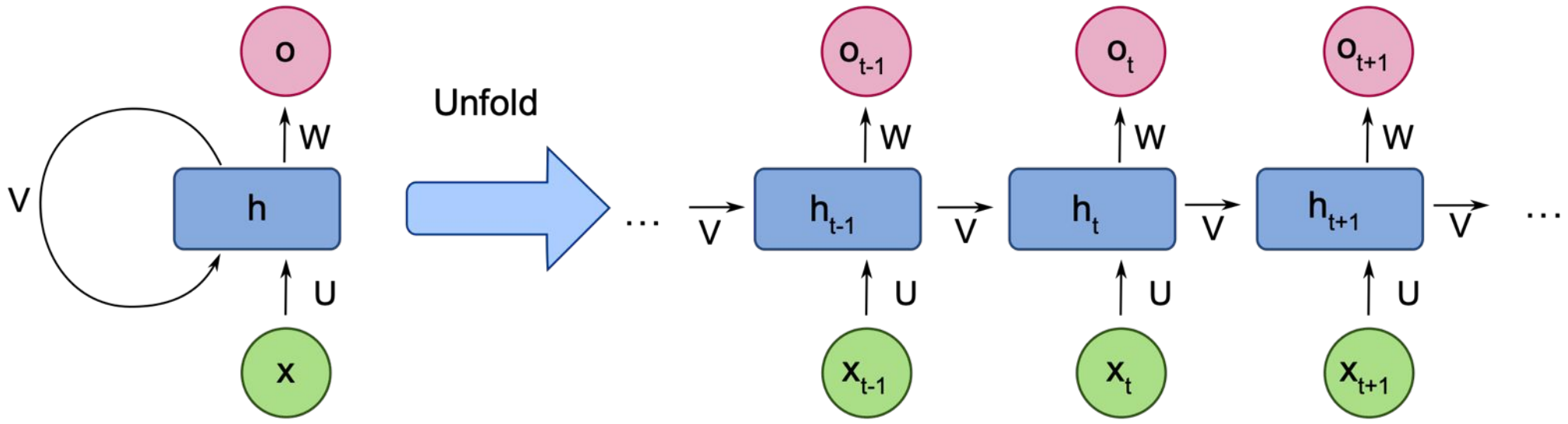
Embedding Layer: Converts characters into dense vectors.

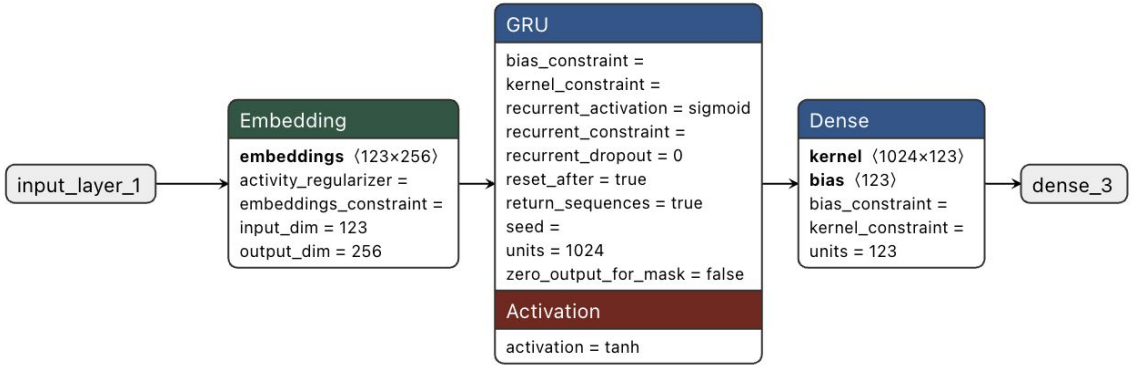
RNN/GRU Layer (1024 units): Learn from sequences.

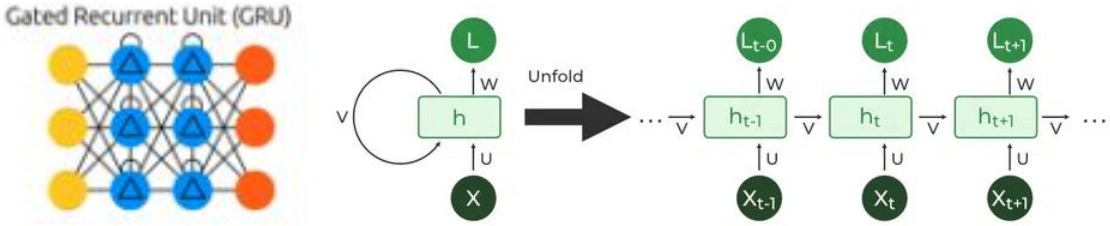
Dense layer: Generates probabilities for each character (123).

Deep Learning models (or artificial neural networks)

Recurrent Neural Networks (RNNs): Designed for **sequential data like time series or text**, these networks use their internal state (memory) to process sequences of inputs.







Model: "sequential_4"

Layer (type)	Output Shape	Param #
embedding_4 (Embedding)	(1, 120, 256)	31,488
gru_3 (GRU)	(1, 120, 1024)	3,938,304
dense_3 (Dense)	(1, 120, 123)	126,075

Total params: 4,095,867 (15.62 MB)
Trainable params: 4,095,867 (15.62 MB)
Non-trainable params: 0 (0.00 B)

RNN MODEL (RECURRENT)

MODEL TRAINING

Loss Function: Categorical Sparse Crossentropy

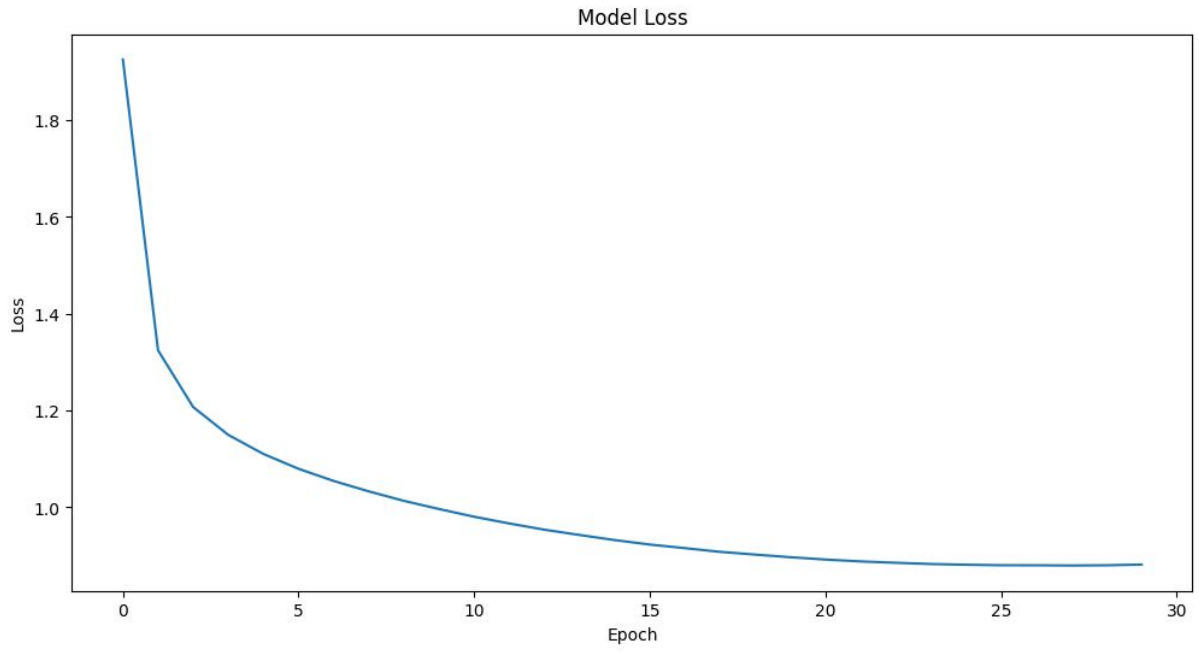
Optimizer: Adam

Epochs: 30

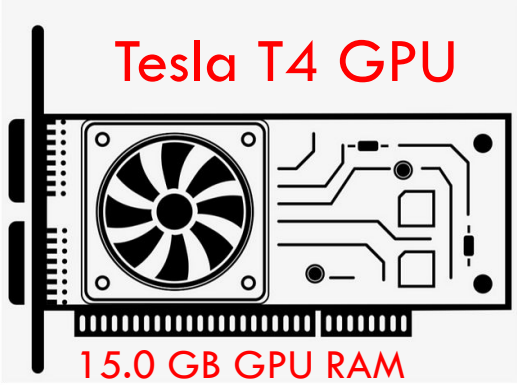
batch size: 128

buffer size: 10,000

Monitoring loss reduction over time.



(33 minutes for training)



TEXT GENERATION

The template generates text character by character from an initial text:

"THE FLYING SUBMARINE".

Temperature controls randomness (0.5 for predictable, 1.0 for creative text).

Generated text with temperature 0.5:

THE FLYING SUBMARINE

CHAPTER 100 VENTANTILE
This eBook is for the use of anyone anywhere in the United States and most other parts of the earth and miserable eruptions. The solar rays should be entirely under the shock of the intensity of the sea. We were all sorts. Are we to prepare for our feelings?"

"I can never see them a good geographer," said Mary.

"Well, then, John, for I get to the Pampas, that we ought to obey the same time. In the country of this latitude changed my brother, and the Nautilus floated in a sea which contained the rudder and lower colour visibly. The loiter was a fatalint region the two scientific discoverers. Several times turning toward the river, the cry of doors and over an inclined plains of the Angara, with a threatening water and disappeared in the midst of the solar rays.

The weather was spread and strewn with closed bottoms which soon appeared that the unexpected sheets of wind was soon and linen, and the whole seas were again landed on the subject of the natives, and the prisoners were successively assuming the sides of this agreement for fifteen days with a threatening voice.

The clouds had disappeared to the ground, and the river and the ship's conditions of the ship's course was still standing, but in his daring explorers, and the sailor thought for the sudden discovery.

"There are no trees, and a half an hour or the sun will soon be carried off they were bringing a special track."

"As you see, my dear Helena, who was the matter of despair?" cried the captain.

"The raft we had done now, captain, and I have a reasonable face of the shipwrecked creek, had been discovered at the entrance of the world."

CHALLENGES AND LIMITATIONS

Limited context window (150 characters).

Difficulty in maintaining coherence in long texts.

Character-level modeling vs. word-level modeling.

CONNECTING WITH MODERN LANGUAGE MODELS

Our Model (VerneBot) :

- **Training data: 5.8 million characters (bytes) (7 books).**
  4 million parameters,
  Character-level tokenization (150)
  RNN architecture.

Open AI GPT-3 (2020):

- **Training data: 45 Trillion bytes (text)**
  175 billion parameters,
  Subword tokenization (2,048 tokens),
  Transformer Architecture.

Modern models handle long-range dependencies better.

SLM

Phi-3-mini
(3.8B)

Phi-3-vision
(4.2B)

Phi-3-small
(7B)

Phi-3-medium
(14B)

- **Architecture: Transformer – 3.8 Billion Parameters**
  Inputs: Text.
  Context length: 128k tokens
  GPU: 512 H100-80G
  Training time: 7 days
  Training data: 3.3 Trillion tokens**

= 150 K tokens

~ 350 pages

~ 300 words/page

1 word = ~ 1.4 token

** Equivalent to 23 million books, that is:
17% of All the books in the world

Questions?

TINYML4D