Natural language generation as dynamical systems

Jiaxi Zhao

13th May, 2024

Language modeling

Language is very different from physics. Although humans began to study both at an early age, the theoretical foundations of language are far behind those of physics. Most successful approaches originate from the statistical modeling of the language, e.g. statistical learning and generative models etc.

To model the language, the first task is to digitalize the text, which is made up of words, into a sequence of numbers that can be manipulated by the computer.

Word2Vec

Word2Vec[1] contains algorithms, e.g. continuous Bag-of-Words and continuous Skip-gram models to learn the embedding vector of each word using corpus data.

vec("Berlin") - vec("Germany") + vec("France") = vec("Paris")

vec("Russian") - vec("River") = vec("Volga River")

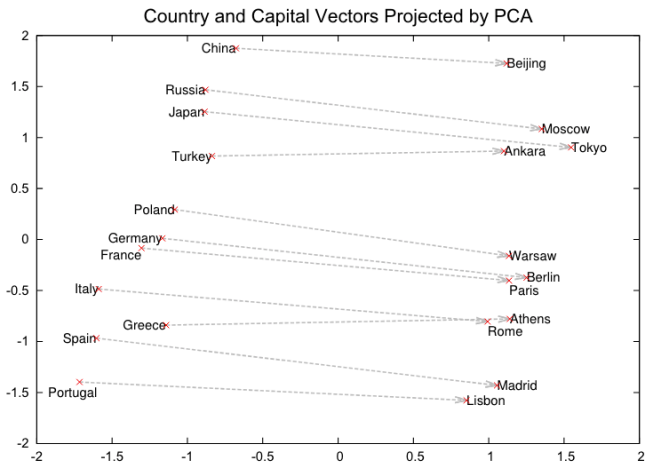


Figure: Word2Vec: PCA from 1000D embedding to 2D

[1] Distributed representations of words and phrases and their compositionality

Seq2Seq

Besides words and phrases, sentences can also have embeddings via the Seq2Seq framework:

$$\text{Input Seq} \overset{\text{Encoder}}{\Longrightarrow} \text{fixed-dim Vec} \overset{\text{Decoder}}{\Longrightarrow} \text{Output Seq.} \quad (1)$$

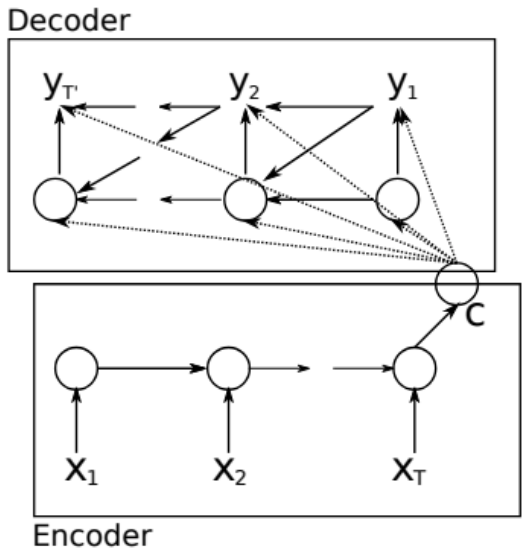


Figure: Seq2Seq

[2]Sequence to Sequence Learning with Neural Networks

Seq2Seq

Sentence embedding also has linguistic meaning:

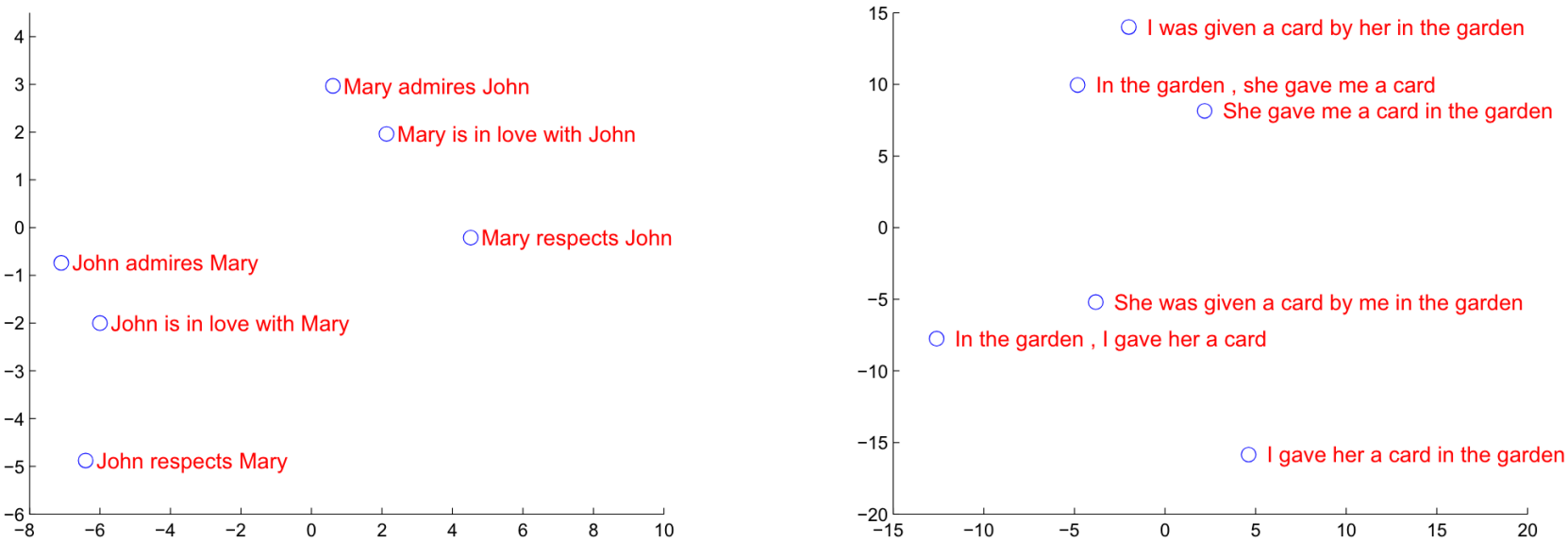


Figure: Seq2Seq

[2]Sequence to Sequence Learning with Neural Networks

Auto-regressive model: Tokenization & embedding

The definition of natural language: *Natural Language Generation, otherwise known as NLG, is to produce natural written or spoken language from structured and unstructured data.*

A step-by-step look at how LLMs generate a much longer sequence given a short prompt.

$$p(x_0, x_1, x_2, ..., x_T) = \prod_{t=1}^{T} p(x_t | x_{<t}). \tag{2}$$

From text to a sequence of vectors:

1. Tokenization: The input text is tokenized into a sequence of tokens.
2. Word embedding: Each token is embedded into a high-dimensional vector space.
3. Positional encoding: The position of each token is encoded into the vector space.

Auto-regressive model: Attention

In self-attention, the query, key, and value are linear transformations of the input X:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V. \tag{3}$$

Notice that this calculation does not mix the vectors (tokens) at different positions. Then, a causal-masked attention is applied to obtain the attention matrix and output:

$$A(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}} \odot M)V. \tag{4}$$

For multi-head attention, we split the query, key, and value into several small parts and concatenate the outputs to obtain the final output

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O, \\ \text{where} \quad \text{head}_i &= A(QW_i^Q, KW_i^Q, VW_i^Q). \end{aligned} \tag{5}$$

[3] Attention is all you need

Auto-regressive model: Linear layer

An MLP is applied immediately after the attention layer to obtain the output:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2. \tag{6}$$

Similar to the query, key, and value projection, this calculation does not mix the vectors (tokens) at different positions.

Dynamics in the autoregressive model

There are at least two kinds of dynamics in the autoregressive model:

- The dynamics of the hidden states in the transformer, The dynamics of the hidden states in the transformer architecture is intensively studied in the NLP literature [4].
- The dynamics of the autoregressive process.

However, as stated by Ilya Sutskever, the co-founder and chief scientist of OpenAI, **Next-token prediction is enough for AGI**. This dynamic is of great importance but not well understood.

[4] How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings

Mistral 7B

We use Mistral 7B[1] as the platform for our experiments, as it is reported to be the most powerful language model for its size up to its release day (Sep 27, 2023). **You really should try it out!**

- Outperforms Llama 2 13B on all benchmarks
- Outperforms Llama 1 34B on many benchmarks
- Approaches CodeLlama 7B performance on code, while remaining good at English tasks
- Uses Grouped-query attention (GQA) for faster inference
- Uses Sliding Window Attention (SWA) to handle longer sequences at smaller cost

[1] https://arxiv.org/pdf/2310.06825, https://huggingface.co/mistralai/Mistral-7B-v0.1, https://mistral.ai/news/announcing-mistral-7b/

Mistral 7B

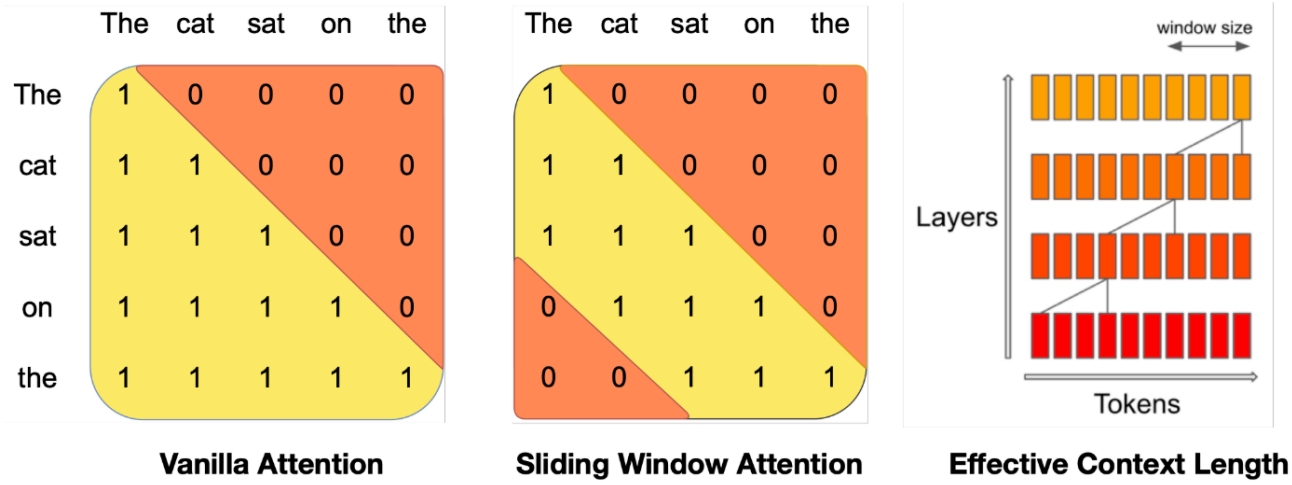


Figure: Sliding Window Attention

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
window_size	4096
context_len	8192

Decoding strategies

1. Greedy decoding

$$x_t = \arg\max_x p(x_t|x_{<t}). \tag{7}$$

2. Beam search
3. Sampling: some out-lier word may spoil the NLG.
4. Top-k sampling: sampling from the first k important tokens, no universal choice of k.
5. Top-p sampling: sampling from the first several important tokens which consist of p of the probability mass.

Decoding strategies

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, b=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Sampling, t=0.9

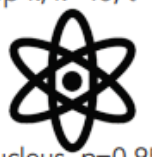
Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Top-k, k=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-k, k=40, t=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.



Nucleus, p=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Visualization of the word vector evolution

Using the prompt *"Long time ago, there was a nymph,"*, we generate a sentence until EOS and visualize the sequences of word vectors over a low-dimensional space.

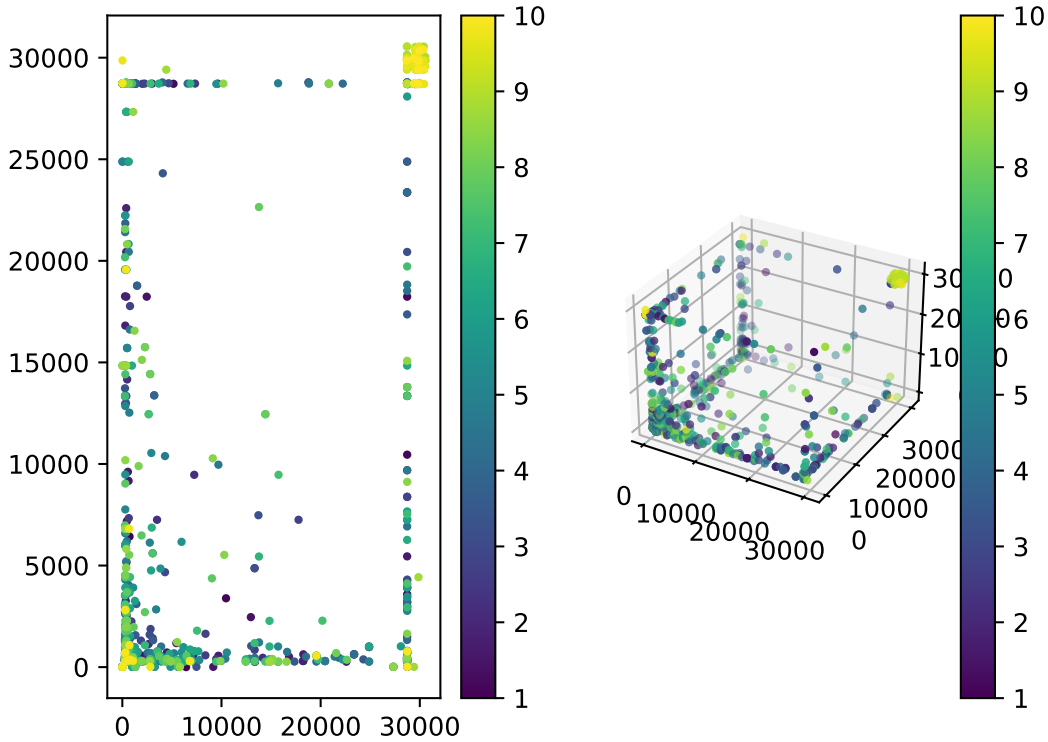


Figure: Delayed embedding

Visualization of the word vector evolution

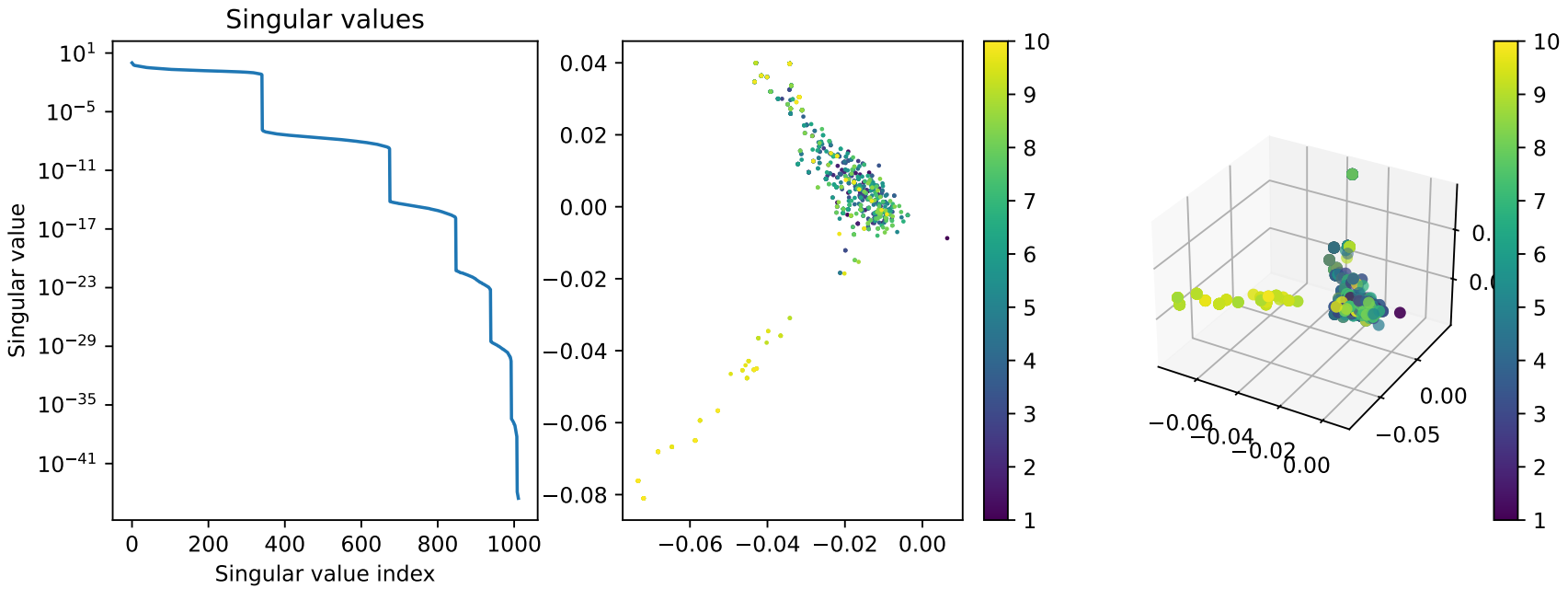
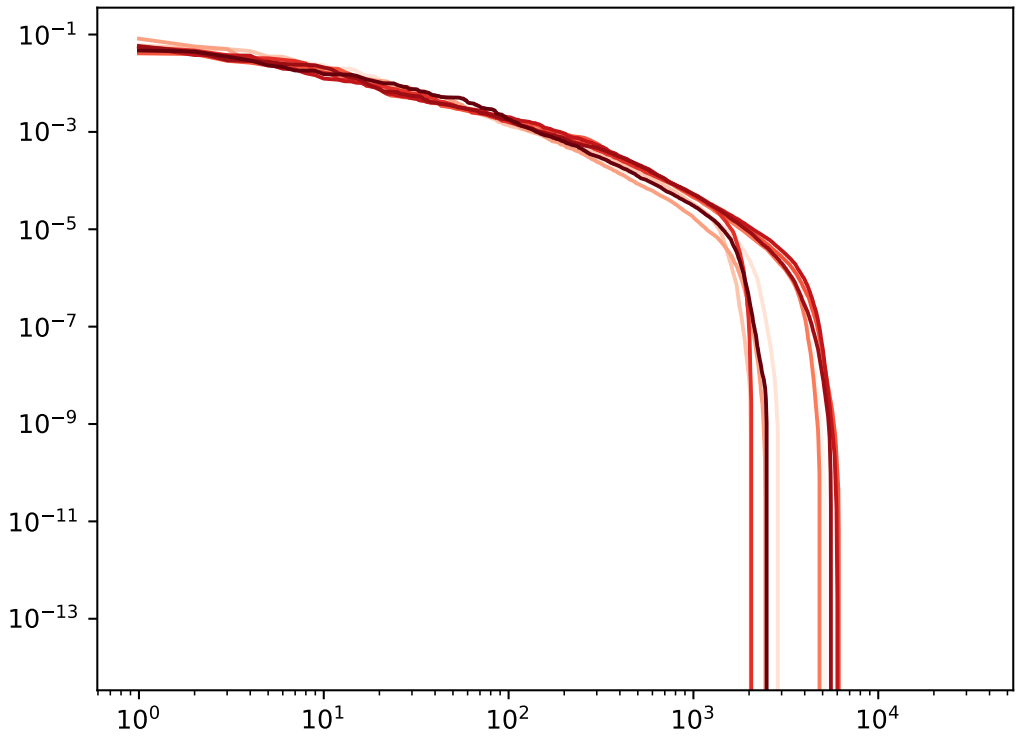


Figure: POD of word vector evolution

Visualization of the probability evolution

The well-known Zipf law states that the frequency of a word is inversely proportional to its rank in the frequency table

$$P(X = k) \propto \frac{1}{k^{\alpha}}. \tag{8}$$



Visualization of the probability evolution

Similar to the idea of Word2Vec, we would like to use this distribution to perform clustering on the vocabulary, which implies their dynamic correlation in NLG. Namely, the probability distribution of shape [vocab size, seq length] can be viewed as another embedding of the word to high-dimensional space

Visualization of the probability evolution

From the other perspective, we can also perform dimension reduction over the vocab_size dimension to visualize the dynamic of the probability evolution:

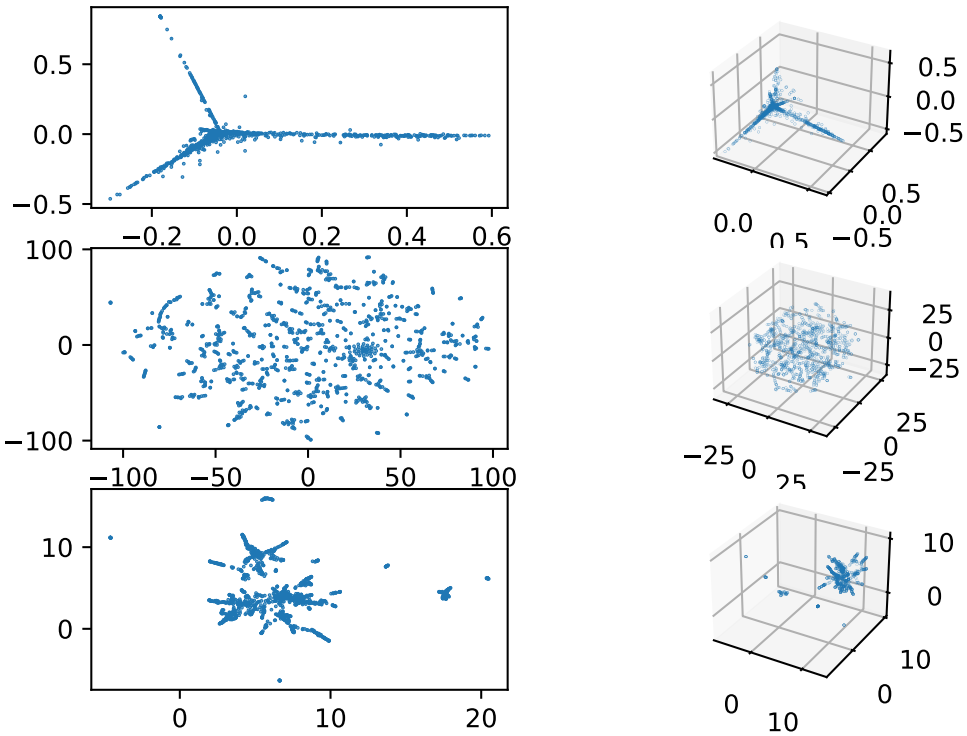
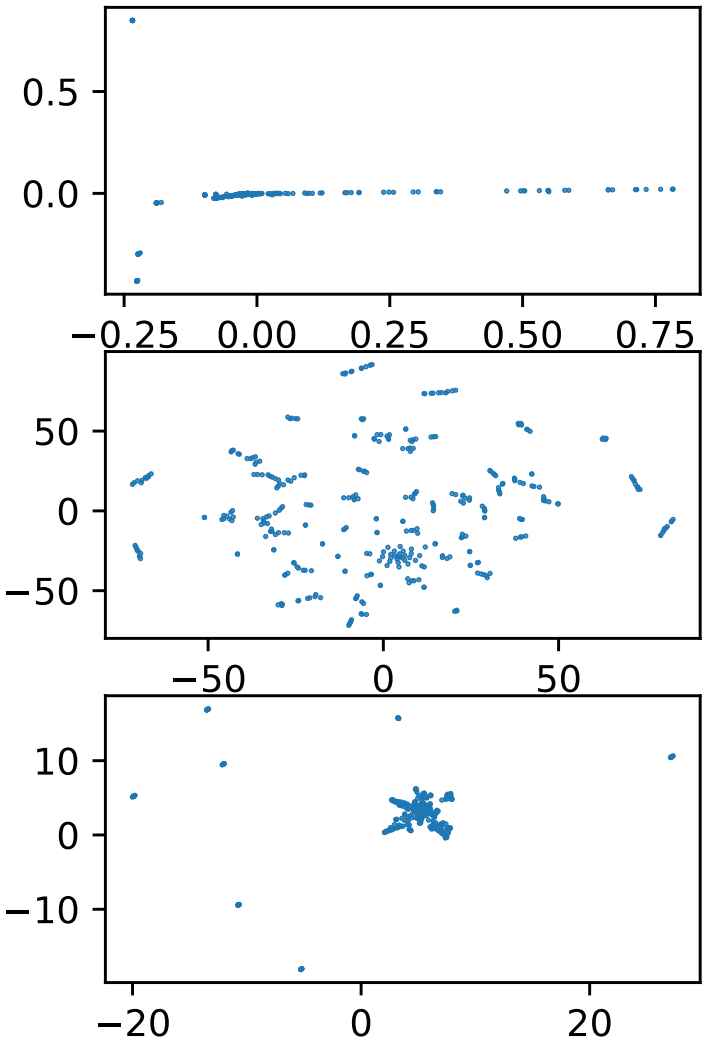
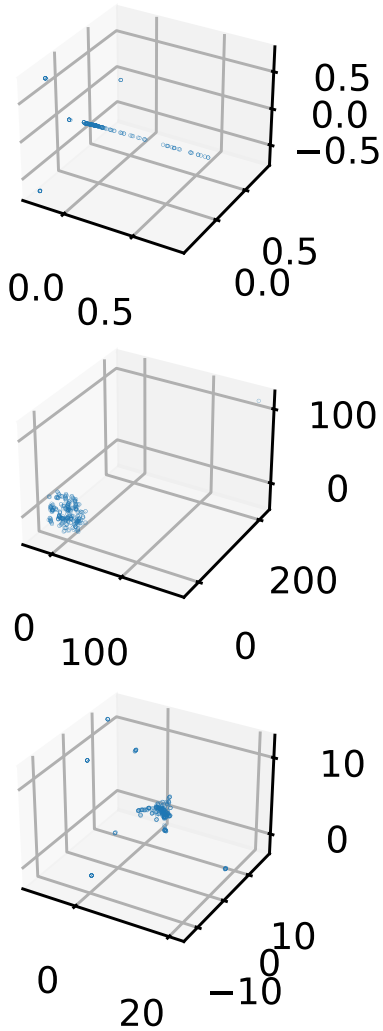


Figure: Low-dim visualization of the distribution evolution

Visualization of the probability evolution

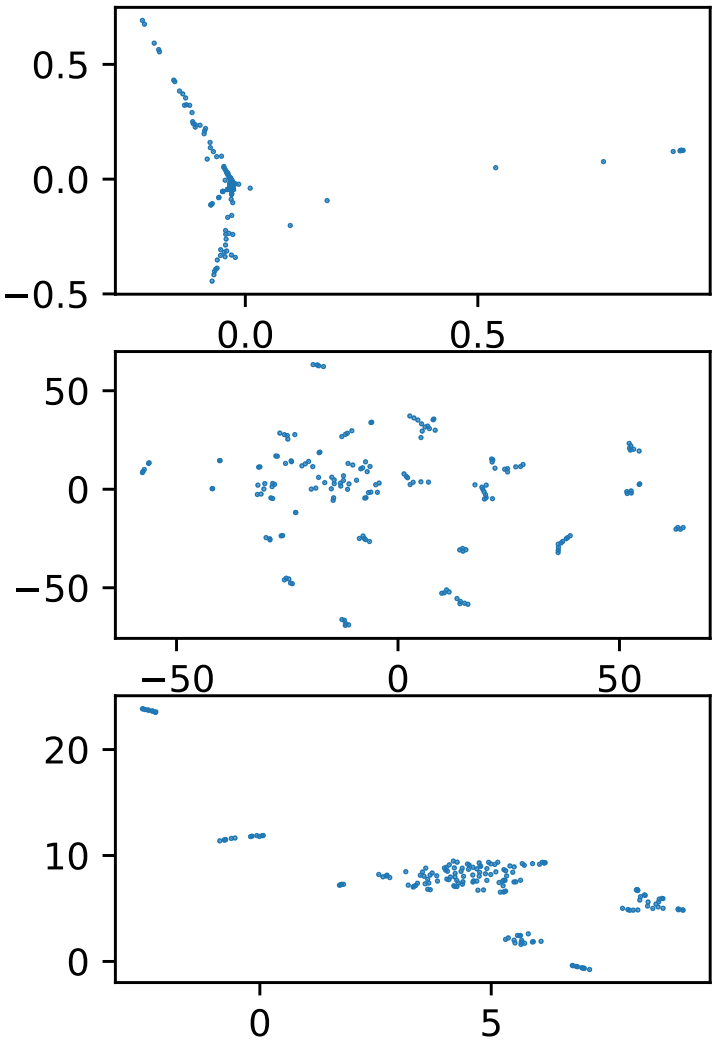
More interesting patterns

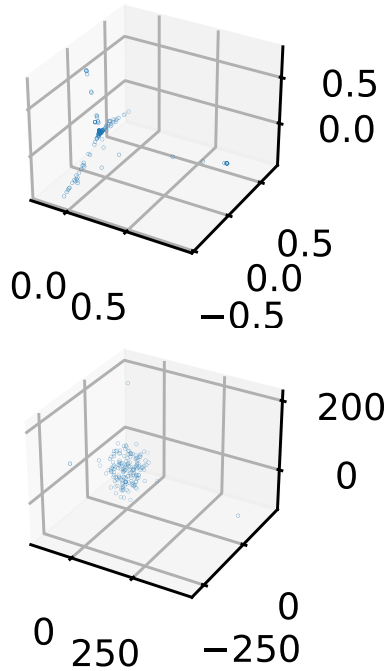




Visualization of the probability evolution

More interesting patterns





State variable of the NLG

Key question: *Viewing the NLG as a dynamical system, what is the state variable of it?*

We need some quantities that are informative enough to encode the whole information in a forward pass of the LLMs. Here are some obvious choices:

1. Zero padding for shorter sequences at the beginning and use the maximum length as the fixed dimension hidden state (usually 4k, even 32k). JX[Markovian model reduction]
2. Use the attention matrices of all heads or their spectral information.
3. In Seq2Seq, the hidden state of the encoder can be used as the state variable. But decoder-only autoregressive models like GPT do not belong to the Seq2Seq category(personal idea), do they?
4. JX[This may be a non-Markov model, need to use another definition of chaoticity.]

Future directions

Further questions:

1. How does prompt engineering change the dynamics of the NLG in LLMs? JX[This seems to be an interesting direction. But the prompt engineering is far from choosing the initial condition, it is more like choosing the boundary so that the generator will not go out of this boundary.]

2. Can we use the tools from information theory such as minimal description length to understand the the dynamics of the NLG?

3. Investigate the dynamics of the autoregressive process in the LLMs for different contexts, e.g. novel, academic journal, and code etc. Ablation studies on different LLMs and different scales should also be explored.

4. What's more?

Reference

1. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

2. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).

3. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing s

4. Jiang, Albert Q., et al. "Mistral 7B." arXiv preprint arXiv:2310.06825 (2023).

Reference

5. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings (Ethayarajh, EMNLP-IJCNLP 2019)

6. Holtzman, Ari, et al. "The curious case of neural text degeneration." arXiv preprint arXiv:1904.09751 (2019).