Week 8 Assignment

Generative AI

Saarland University – Winter Semester 2024/25

Martínez
7057573
cama00005@stud.uni-saarland.de

1 Exercise Assignment: E1

Algorithm 1: Text Generation with Hard Red List

Input: prompt, $s^{(-N_p)}, \ldots, s^{(-1)}$

for $t = 0, 1, \ldots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)}, \ldots, s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this seed, randomly partition the vocabulary into a "green list" G and a "red list" R of equal size.
4. Sample $s^{(t)}$ from G, never generating any token in the red list.

Algorithm 2: Text Generation with Soft Red List

Input : prompt, $s^{(-N_p)} \ldots s^{(-1)}$
green list size, $\gamma \in (0, 1)$
hardness parameter, $\delta > 0$

for $t = 0, 1, \ldots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \ldots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this random number generator, randomly partition the vocabulary into a "green list" G of size $\gamma|V|$, and a "red list" R of size $(1-\gamma)|V|$.
4. Add δ to each green list logit. Apply the soft-max operator to these modified logits to get a probability distribution over the vocabulary.
5. $$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)}+\delta)}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in R. \end{cases}$$
6. Sample the next token, $s^{(t)}$, using the watermarked distribution $\hat{p}^{(t)}$.

This exercise assignment aims to give you a deeper understanding of the watermarking of LLMs as described in the paper *A Watermark for Large Language Models* [1].

More concretely, we consider a simplified vocabulary $V = \{t_0, t_1, t_2, t_3, t_4\}$, where t_0 is a special token that marks the beginning of a sentence.

(E.1) First, consider Algorithm 1 in [1], *Text Generation with Hard Red List*. The possible tokens in the Red List depend on the previous token, as shown in Table 1. Based on the presence of

Table 1: Red List for text generation

x_{i-1}	Tokens in Red List
t_0	t_1, t_2
t_1	t_1, t_2
t_2	t_2, t_3
t_3	t_1, t_4
t_4	t_3, t_4

any red token in the sentence, which of the following sentences contains a watermark, and which does not?

- $t_0t_1t_2t_3t_4t_1t_4$
- $t_0t_4t_1t_3t_3t_2t_1$

(E.2) Consider the text $t_0t_3t_2t_4t_1t_3t_3t_3$, which was generated by a watermarked model. How could you remove the watermark by inserting a single new token? What would the resulting text look like?

(E.3) Consider the text $t_0t_1t_3t_2t_1t_4t_2$, which was generated by a non-watermarked model. How could you make this text look like it was generated by a watermarked model by inserting a single new token? What would the resulting text look like?

(E.4) Consider a model that predicts each token with equal probability, with all the logit values $l^{(i)}_{t_1} = l^{(i)}_{t_2} = \ldots = 1$ at any time step i. What would be the probability that the model generates the following sentences when not considering any watermarking?

- $t_0t_1t_4t_3t_4t_1t_4$
- $t_0t_4t_1t_3t_3t_2t_1$

Remark: Note that the model generates tokens one by one in order to generate a sentence (refer to week 2 slides). Here, t_0 marks the beginning of the sentence and you can set the probability of the first token being t_0 as 1.0.

(E.5) Now, consider Algorithm 2 in [1], *Text Generation with Soft Red List*. Use the same Red List as in Table 1. When using $\delta = 1$ and the same model as in the previous exercise, what is the probability that the following sentences are generated considering the Soft Red List watermarking method?

- $t_0t_1t_4t_3t_4t_1t_4$
- $t_0t_4t_1t_3t_3t_2t_1$

2 Exercise Assignment: E2

3 Exercise Assignment: E3

4 Exercise Assignment: E4

5 Exercise Assignment: E5

6 Exercise Assignment: I1

7 Exercise Assignment: I2

8 Exercise Assignment: I3

9 Exercise Assignment: I4

10 Exercise Assignment: I5

Acknowledgements

This week's slides and listed references.

References