Protecting Language Models via Invisible Watermarking

> https://arxiv.org/pdf/2302.03162.pdf

- 这篇文章提供了一个描述watermark动机的全新视角：除了减少大模型滥用，也可以一定程度避免模型蒸馏出低价替代品；因为这个水印是加在大模型的decode过程中的，一个well-trained的知识蒸馏出来的也会带着水印，这样只要对声称是自研模型的低价替代品生成的文本做水印检测，如果检测出是带有某个版权LLM的水印，就可知是从哪蒸馏出来的。
- GINSEW：Generative Invisible Sequence Watermarking
  - Problem Setting：需要一个可信的检测方，要有对suspect model的white-box access，以及一个测试数据集，用一个key去对比这俩模型的输出，看它们的secret signal是否match

    感觉这个检测成本比较大，还需要white-box access；
  - Watermarking Process：全局红绿列表 + 伪随机数

Algorithm 1 Watermarking process

Inputs: Input text $\boldsymbol{x}$, probability vector $\mathbf{p}$ from the decoder of the victim model, vocab $\mathcal{V}$, group 1 $\mathcal{G}_1$, group 2 $\mathcal{G}_2$, hash function $g(\boldsymbol{x}, \mathbf{v}, \mathbf{M})$.
Output: Modified probability vector $\mathbf{p}$
Calculate probability summation of tokens in group 1 and group 2: $Q_{\mathcal{G}_1} = \sum_{i \in \mathcal{G}_1} \mathbf{p}_i$, $Q_{\mathcal{G}_2} = \sum_{i \in \mathcal{G}_2} \mathbf{p}_i$
Calculate the periodic signal

$$z_1(\boldsymbol{x}) = \cos\left(f_w g(\boldsymbol{x}, \mathbf{v}, \mathbf{M})\right),$$
$$z_2(\boldsymbol{x}) = \cos\left(f_w g(\boldsymbol{x}, \mathbf{v}, \mathbf{M}) + \pi\right)$$

Set $\tilde{Q}_{\mathcal{G}_1} = \frac{Q_{\mathcal{G}_1} + \varepsilon(1+z_1(\boldsymbol{x}))}{1+2\varepsilon}$, $\tilde{Q}_{\mathcal{G}_2} = \frac{Q_{\mathcal{G}_2} + \varepsilon(1+z_2(\boldsymbol{x}))}{1+2\varepsilon}$
for $i = 1$ **to** $|\mathcal{V}|$ **do**
  if $i \in \mathcal{G}_1$ **then** $\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_1}}{Q_{\mathcal{G}_1}} \cdot \mathbf{p}_i$
  else $\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_2}}{Q_{\mathcal{G}_2}} \cdot \mathbf{p}_i$
end for
return $\mathbf{p}$

    - 全局红绿列表：g1, g2
    - 对新token的logits引入随机数signal
  - Watermark Detection

Algorithm 2 Watermark detection

Inputs: Suspect model $\mathcal{S}$, sample probing data $\mathcal{D}$ from the training data of $\mathcal{S}$, vocab $\mathcal{V}$, group 1 $\mathcal{G}_1$, group 2 $\mathcal{G}_2$, hash function $g(\boldsymbol{x}, \mathbf{v}, \mathbf{M})$, filtering threshold value $q_{\min}$.
Output: Signal strength
Initialize $\mathcal{H} = \emptyset$
for each input $\boldsymbol{x}$ in $\mathcal{D}$ **do**
  $t = g(\mathbf{v}, \boldsymbol{x}, \mathbf{M})$
  for each decoding step of $\mathcal{S}(\boldsymbol{x})$ **do**
    Get probability vector $\hat{\mathbf{p}}$ from the decoder of the suspect model.
    $\hat{Q}_{\mathcal{G}_1} = \sum_{i \in \mathcal{G}_1} \hat{\mathbf{p}}_i$
    $\mathcal{H} \leftarrow \mathcal{H} \cup (t, \hat{Q}_{\mathcal{G}_1})$
  end for
end for
Filter out elements in $\mathcal{H}$ where $\hat{Q}_{\mathcal{G}_1} \leq q_{\min}$, remaining pairs form the set $\widetilde{\mathcal{H}}$.
Compute the Lomb-Scargle periodogram from the pairs $(t^{(k)}, \hat{Q}^{(k)}_{\mathcal{G}_1}) \in \widetilde{\mathcal{H}}$
Compute P_{snr} in Equation 5.
return P_{snr}

    - **shared key：hash function g, v, M**
    - **access：**
        - **watermarking process: white-box access to victim model**
        - **watermark detection process: white-box access to suspect model**
- Evaluation
    - Task: Machine Translation & Story Generation
    - Model collection: 对每个task，训了一个Transformer base model作为victim model，再根据这个训了20个suspect model作为positive examples，以及30个models作为negative examples
    - Text Quality Evaluation
        - Machine Translation：BLEU & BERTScore
        - Story Generation：ROUGE & BERTScore
    - Detection mAP
    - Robustness Evaluation
        - Watermark removal attack: synonym randomization
- Watermark detection with text alone：效果一般

Algorithm 3 Watermark detection with text alone

Inputs: Suspect model $\mathcal{S}$, sample probing data $\mathcal{D}$ from the training data of $\mathcal{S}$, vocab $\mathcal{V}$, group 1 $\mathcal{G}_1$, group 2 $\mathcal{G}_2$, hash function $g(\boldsymbol{x}, \mathbf{v}, \mathbf{M})$.
Output: Signal strength
Initialize $\mathcal{H} = \emptyset$
for each input $\boldsymbol{x}$ in $\mathcal{D}$ **do**
    $t = g(\mathbf{v}, \boldsymbol{x}, \mathbf{M})$
    $\boldsymbol{y} \leftarrow \mathcal{S}(\boldsymbol{x})$
    for each token of $\boldsymbol{y}$ **do**
        $\mathcal{H} \leftarrow \mathcal{H} \cup (t, \mathbf{1}(\boldsymbol{y}_i \in \mathcal{G}_1))$
    end for
end for
Compute the Lomb-Scargle periodogram from $\mathcal{H}$, and compute P_{snr} in Equation 5.
return P_{snr}