Watermarking for Generative Models (Lecture 6)

PROBLEM: How to detect AI-generated images/text/video?

- Detecting AI-generated misinformation
- Preventing Model collapse ("AI slop")
  - training LLMs on LLM-generated data ⇒ quality loss
- Detecting inappropriate use eg. plagiarism

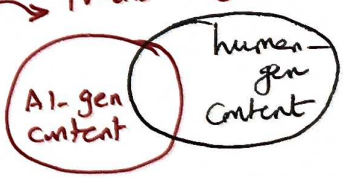
– Many companies that deal with user-gen content care a lot (eg. Youtube, X, Spotify)

APPROACH 1: Post-hoc detection

E.g. GPTZero, DetectGPT, the "em-dash effect".

Hypothesis: Detectable, intrinsic, markers of AI-generated content exist.

↳ TV-dist large?



Problems: False negatives, easy to evade

False positives
- accuse the innocent
- accuse non-native English writers

⊖ Increasingly unrealistic as AI models become more powerful.

Intrinsic markers of AI gen content, if they exist, are generally subtle.

⊕ Do not need model owners to cooperate in detection.

The watermarking approach:

- Intentionally embed markers
- Embed using a secret rule (e.g. using a secret key)
- Typically, detection uses the same key.
  (although public detection may be possible
  e.g. Fairoze et al. Eprint 2023/1661)

Classical Topic in Cryptography & Security

e.g. watermarking images, text, videos etc.

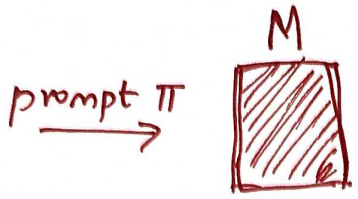
Very hard. OTOH need to change the text to embed a marker; on the other hand, how do I know that my change hasn't degraded quality?

The Key Difference for us:

We have access not only to an image but the generative process that produced the image.

Gives us Lots of freedom & we will exploit it.

Generative Models



we will treat this largely as a black box.

notation: $\text{Model}^M \Rightarrow D_\pi$

e.g. LLMs (autoregressive), Diffusion Models, ...

In fact, today focus on

Autoregressive Models: sample a string from T^*

T = set of tokens (for GPT-4 $|T| \sim 10^5$)

Code for Model M:

```
t = 0, X_0 = ⊥
while X_t ≠ done
    P_{t+1} = M(π, X_1, .., X_t)
    sample X_{t+1} ~ P_{t+1}
    t = t + 1
output X_1, .., X_t.
```

P_{t+1}: probability distribution over T

$M(\pi, X_1, .., X_t)$: neural network model

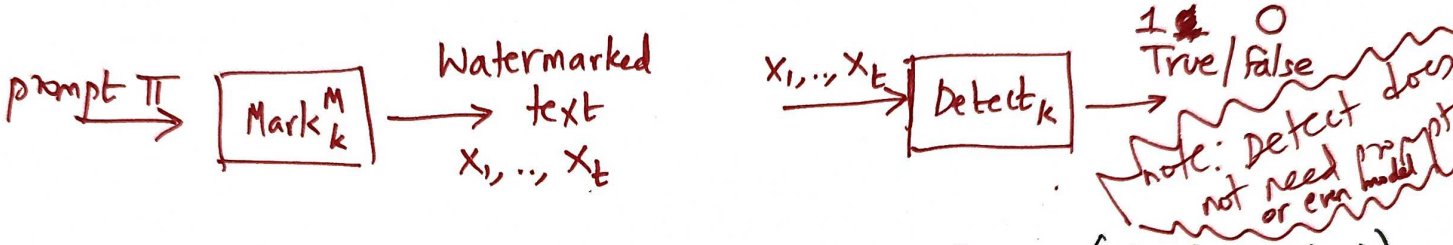
e.g: $M(X_1, .., X_6)$ = "As a language model, I"

P_7 =
- "can": 10%
- "cannot": 70%
- "am": 5%
- "do": 3%
- ...

Watermarking LLMs (= autoregressive models)

The "API": Two algorithms Mark_k^M, Detect_k:

- Mark and Detect share a private key k
- Mark knows the (autoregressive) model M



TBD: What properties do we want from (Mark, Detect)?

Strawman 1:

Recall Pseudorandom functions $\mathcal{F} = \{F_k : \{0,1\}^n \to \{0,1\}^m\}$ s.t. for all p.p.t. algorithms A,

Exercise: key $\geq \lambda$ bits long

$$\left| \Pr[A^{F_k} \to 1] - \Pr[A^{R} \to 1] \right| \leq \text{negl}(\lambda) \cdot 2^{-\lambda}$$

where λ: security parameter
R: uniformly random function

PRFs give us a good (comp. secure) secret-key encryption scheme:

$$\text{Enc}_k(m; r) = (r, PRF_k(r) \oplus m)$$

$$\text{Dec}_k(r, c) = c \oplus PRF_k(r)$$

PRFs also give us a good message authentication scheme (MAC).

$$MAC_k(m) = PRF_k(m)$$

$$Verify(m,t) = (YES \text{ iff } t = PRF_k(m))$$

Security (unforgeability): given $(m_1,t_1),\dots,(m_k,t_k)$, computationally hard for any p.p.t. algorithm to output (m^*,t^*) where (a) $m^* \notin \{m_1,\dots,m_k\}$ AND (b) $t^* = MAC_k(m^*)$

$Wat_k^M(\pi)$: generate $x \sim M(\pi)$
output $(x, MAC_k(x))$

- watermarked texts are correctly recognized (completeness)
- for any given text x and tag t, the chance that $t = MAC_k(x)$ "accidentally" (i.e. for a random k) is small ($\leq 2^{-k}$). (soundness / no false positives)

✗ trivially removable.

Lesson: Watermarks cannot be trivially appended. Have to be intrinsic.

Strawman 2: the red/green scheme

- Randomly partition the tokens into RED and GREEN tokens.

  (Apple, Alphabet, Arugula, Kale, Bacon, Cheese)

- Increase the probability of the words in the green list.

  (in an extreme, only output green tokens)

- Detection is easy (e.g. count # green tokens)

Problems?

1) Quality Degrades: Model prefers not to talk about [redacted] apples.

2) If you talk about bacon or cheese a lot, you may be falsely accused: no soundness.

3) When is the watermark detectable?

   completeness, assuming entropy

   if output $|\pi$ [redacted] is deterministic,

   (the quick brown fox .....)

   no hope to watermark: either change the output & lose quality or don't change the output & lose soundness.

Property 1: Soundness (no false positives) — important! absolute!

for all text t, $\Pr_{sk}\left[\text{Detect}_{sk}(t) = \text{TRUE}\right] \leq 2^{-\lambda}$.

we won't think about this too much

Stronger: unforgeability

Given watermarked $t_1, t_2, \ldots, t_k$ cannot produce $t^* \notin \{t_1, \ldots, t_k\}$ s.t $\text{Detect}_{sk}(t^*) = \text{TRUE}$

Property 2: Completeness ((b-completeness) — conditional on enough entropy

Text generated by watermarked model will be recognized (given sk).

For all prompts Π

$$\Pr_{\substack{sk \\ t \leftarrow \text{Mark}^{M}_{sk}(\Pi)}}\left[\text{Detect}_{sk}(t) = \text{False AND } \underbrace{\text{Entropy}(\Pi) \geq b}_{\text{we will need to refine}}\right] \leq 2^{-\lambda}.$$

Stronger: robustness (w.r.t. modifications $\mathcal{E}$)

..... even after modified using transformation $\mathcal{E}$

A very difficult property to achive in general

we will in next lecture look at substitutions, edits...

For all prompts Π

$$\Pr_{\substack{sk \\ t \leftarrow \text{Mark}^{M}_{sk}(\Pi)}}\left[\text{Detect}_{sk}(\mathcal{E}(t)) = \text{False AND } \text{Entropy}(\Pi) \geq b\right] \leq 2^{-\lambda}.$$

Property: "Quality"

Very hard to define but what if output of watermarked model is "as good as" the unwatermarked model for all (poly time) applications?

↳ Undetectability

For all models &
For all efficient algorithms A,

$$\left| \Pr\left[A^{\text{Model}^M} \to 1\right] - \Pr\left[A^{\text{Mark}^M_{sk}} \to 1\right] \right| \leq \vec{\lambda}.$$

Weaker notion: "distortion-freeness"

For all prompts π and models M,

$$\text{Mark}^M_{sk} \equiv \text{Model}^M(\pi)$$

"Watermarked dist. same as original distribution for a single sample".

.... next page

An undetectable watermarking scheme
(Christ-Gunn-Zamir 2024)

- For simplicity & w.l.o.g. think of token space = $\{0,1\}$
- so, given a ~~prefix~~ prompt π & prefix $x_1, .., x_{t-1}$, the model defines a probability p_t s.t

$$x_t \sim \text{Binomial}(p_t)$$

$$(\Pr(x_t = 1) = p_t)$$

Scheme:

IDEA
Correlated sampling

Key $k = (u_1, u_2, \ldots, u_T) \in [0,1]^T$ (T random real numbers in $[0,1]$)

Mark_k^M: for each t, output $x_t =$ 1 if $u_t \le p_t$; 0 otherwise

- Marginal dist of x_t is Binomial(p_t) indeed. → PERFECT QUALITY
- However, x_t and u_t are correlated. (non-0/non-1)
  (for a fixed p_t, the smaller u_t, the more likely x_t is 1)
- OTOH, if $p_t = 0$ or $p_t = 1$, u_t has no effect on x_t. (no entropy) → COMPLETENESS & SOUNDNESS

$\text{Detect}_k(x_1, .., x_T)$: computes a score $s(x_j, u_j)$ where $u_1, .., u_T$

$$s(x_j, u_j) = \begin{cases} -\ln u_j & \text{if } x_j = 1 \\ -\ln(1-u_j) & \text{if } x_j = 0 \end{cases}$$

$$c(x) = \sum_{j=1}^{T} s(x_j, u_j)$$

if $c(x)$ is large, output watermarked.
else output non-watermarked.

QUICK CALCULATION

Completeness & Soundness: for any fixed text $x_1, \ldots, x_T$ (not correlated with $u_1, \ldots u_T$)

each $s(x_j, u_j)$ is an exponential random var.

$\left(-\log u_{[0,1]} \right)$
pdf(x) $= e^{-x}$ $x \geq 0$

$$\mathop{\mathbb{E}}_{u_j}\left[s(x_j, u_j)\right] = \int_0^1 \ln(1/u)\, du = 1$$

$$\Rightarrow \mathop{\mathbb{E}}_{u_1 \ldots u_T}\left[c(x) - \underbrace{|x|}_{=T}\right] = 0$$

for watermarked outputs,

$$\mathop{\mathbb{E}}_{u_j}\left[s(x_j, u_j)\right] = \int_0^{p_j} \ln\left(\frac{1}{u}\right) du$$

$$\int_0^1 s(x(u), u)\, du = \int_0^1 \mathbb{1}[u \leq p_j] \cdot \ln\left(\frac{1}{u}\right) + \mathbb{1}[u > p_j] \cdot \ln\left(\frac{1}{1-u}\right) du$$

$$= \int_0^{p_j} \ln\left(\frac{1}{u}\right) du + \int_{p_j}^1 \ln\left(\frac{1}{1-u}\right) du$$

$$= \int_0^{p_j} \ln\left(\frac{1}{u}\right) du + \int_0^{1-p_j} \ln\left(\frac{1}{u}\right) du$$

(11)

$$= p_j - p_j \ln p_j + (1-p_j) - (1-p_j)\ln(1-p_j)$$

$$= 1 + \ln(2) \cdot H(p_j)$$

$\int \ln u = u \ln u - u$

Detect(π)
if $c(x) - |x| \geq \lambda\sqrt{|x|}$ say "watermarked"
else "not watermarked"

So,

$$\mathop{\mathbb{E}}_{u_1,\ldots u_T}\left[c(x) - |x|\right] = \ln 2 \cdot H(\underbrace{\text{Model}(\pi)}_{\text{distribution on input prompt } \pi})$$

Problem 1 Expectations different but doesn't guarantee Completeness & soundness

Need to analyze e.g. variance.

(what if high Shannon entropy but outputs a fixed $(x_1,\ldots,x_t)$ with prob $1/2$?)

(is the $1/2$ prob output watermarked? = output by Mark_k^M & detected by Detect_k?
If so, no soundness.
If not, no ~~completeness~~ Undetectability or no completeness.) if model decides not to output it

Way out: require completeness only for outputs with high empirical entropy

$$\text{Emp}\cdot H^{M,\pi}(x) = -\log \Pr\left[\text{Model}^M(\pi) = x\right]$$

$$H(\pi) = \mathop{\mathbb{E}}_{x \sim \text{Model}(\pi)}\left[\text{Emp}\cdot H^{M,\pi}(x)\right]$$

Problem 2 Long keys

Use pseudorandomness

key = k $\quad u_j = prf_k(j)$

Problem 3 multiple outputs

Can't use the same u_j → not undetectable

(once you fix $(u_1, \ldots, u_T)$, model is deterministic!)

how about $u_j = prf_k(\pi \| j)$ → detector doesn't know π & what if π repeats?

how about pick random nonce r

and let $u_j = prf_k(r \| j)$

Concatenate r to model output → same issue as MAC

Solution: generate the first ℓ tokens until they have accumulated enough entropy

i.e. until $\sum_{i=1}^{\ell} -\log \Pr(x_i \mid x_1, \ldots, x_{i-1}) \geq \lambda$.

$\quad$ " $-\log \Pr(x_1, \ldots, x_\ell)$

- use $x_1, \ldots, x_\ell$ as "seed".
- set $u_j = prf_k(x_1, \ldots, x_\ell \| j)$

— Detect does not know ℓ so it tries all possibilities ($\ell = 1, 2, 3, \ldots, T$)

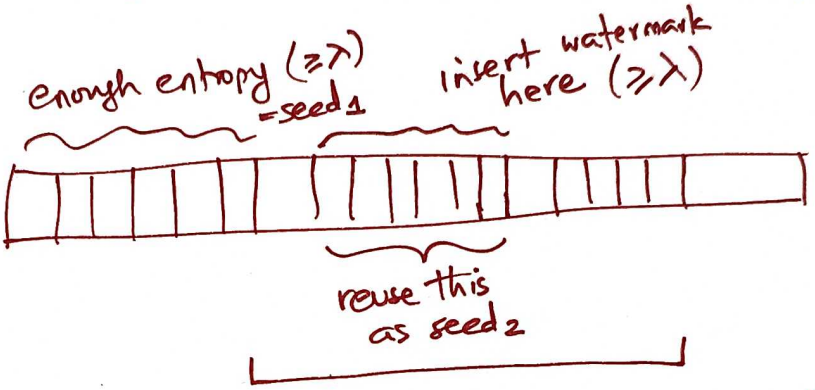
problem 3) unremovability ?

:(easily removed if ~~one~~ many of the seed bits is changed

a fix: substring-robustness:

verification only needs a sufficiently long contiguous substring

Idea:



as long as some long enough contiguous shj is observed, watermark is detectable.

Typos from here on!!! → Refer to CGZ for formal proofs!

Additional Material: full construction + proof sketch (without substring-robustness)

$\text{Mark}_k^M(\pi)$

$i = 1$

$H = 0$

While done $\notin (x_1, \ldots, x_{i-1})$ do:

$p_i = \text{Model}(\pi, x_1, \ldots, x_{i-1})$;

if $H < \lambda$ // if not enough entropy yet

  sample $x_i \leftarrow p_i$ // collect more entropy

  $H = H + \left(-\log p_i \cdot x_i + (1-p_i)(1-x_i)\right)$ // how much add'l entropy did I accumulate

  if $H \geq \lambda$ then

    $r = (x_1, \ldots, x_i)$ // r is the input to the PRF

else

  $x_i \leftarrow 1$ if $F_{sk}(r, i) \leq p_i$

    0 o/w // embed watermark

end.

$i = i+1$

end.

$\text{Detect}_k(x_1, \ldots, x_T)$:

so enumerate over all choices

don't know how long PRF input is

for $i \in [T]$ do

$r^{(i)} = (x_1, \ldots, x_i)$

Define

$$V_j^{(i)} = \begin{cases} F_{sk}(r^{(i)}, j) & \text{if } x_j = 1 \\ 1 - F_{sk}(r^{(i)}, j) & \text{if } x_j = 0 \end{cases}$$

if

$$\sum_{j=i+1}^{L} \ln\left(1/V_j^{(i)}\right) > (L-i) + \lambda\sqrt{L-i}$$

return true

end.

if score is high enough, decide "text is watermarked".

For analysis, assume F_{sk} is a random function, invoking the pseudorandomness guarantee + fact that only black box use of F_{sk} + violation of completeness/soundness/undetectability is testable.

Soundness Fix some text $x_1, \ldots, x_T$ (ind of sk).

Want: $\Pr[\text{Detect}(x) = \text{TRUE}] \leq 2^{-\lambda}$.

O is a random fn.

For $i, j \in [T]$ define (1) $r^{(i)}$

(2) $u_j^{(i)} = O(r^{(i)}, j)$

$$v_j^{(i)} = \begin{cases} u_j^{(i)} & \text{if } x_j = 1 \\ 1 - u_j^{(i)} & \text{if } x_j = 0 \end{cases}$$

$u_j^{(i)}$ is independent of x_j so

$v_j^{(i)} \sim U([0,1])$

$E_j^{(i)} = \ln(1/v_j^{(i)})$ are exponential random variables w/ rate 1.

Fix i: $\Pr\left[\sum_{j=i+1}^{L} E_j^{(i)} > (L-i) + \lambda\sqrt{L-i}\right] \leq 2^{-O(\lambda)}$.

Union bound over $i \in [T]$: $\Pr[\exists i \in [T] \text{ st } !] \leq T \cdot 2^{-O(\lambda)}$

$= \text{negligible in } \lambda$.

Undetectability

~~Wat$^0_M(\Pi)$~~

Given a fixed history of $(\text{prompt}_i, \text{response}_i)_{i<\ell}$ pairs, we will show that $\text{Wat}^0_M(\Pi)$ is stat. indistinguishable from $\text{Model}^M(\Pi)$.

if $r_i = \text{prefix}(x_i)$ is distinct from

$r = \text{prefix}(\text{Model}^M(\Pi))$

then $O(r, \star)$ is an independent random function from

$O(r_i, \star)$ for all i

Therefore $\text{Wat}^0_M(\Pi) \equiv \text{Model}^M(\Pi)$

All that remains is to bound collision probability.

Since r has empirical entropy $\geq \lambda$ by def, this is

at most $T \cdot 2^{-\lambda} = \text{negl}(\lambda)$. ∎

Completeness

Want:
$$\Pr_{\substack{O \\ x \leftarrow \text{Wat}^0(\Pi)}} \left[\text{Detect}^O(x) = \text{false AND } H_e(M, \Pi, x) \geq 8\lambda\sqrt{L} \right] \leq 2^{-\lambda}.$$

will show a stronger statement: for ANY $x \in T^*$ and Π st $H_e(M, \Pi, x) \geq 8\lambda\sqrt{L}$

~~& each bit of x has empir.~~

$$\Pr_O \left[\text{Detect}^O(x) = \text{false} \mid \text{Wat}^O(\Pi) = x \right] \geq 2^{-\lambda}.$$

r.v. $S_j = \ln\left(\frac{1}{V_j}\right)$ Conditioned on $\text{Mark}^0(\pi) = x$ or eq. on x_j.

If $x_j = 1$

$S_j = \ln\left(\frac{1}{U_j}\right)$ conditioned on $U_j \le P_j$

U_j → random in $[0,1]$

$U_j \le P_j$ → random in $[0, P_j]$

So S_j is exponential conditioned on $S_j \ge \ln \frac{1}{P_j}$

Fact exponential r.v.s are memoryless

So the conditional dist of S_j cond. on $x_j = 1$ is $\ln\left(\frac{1}{P_j}\right) + E_j$ where E_j is exponential with rate 1.

Similarly for $x_j = 0$

.....

the conditional dist of S_j cond. on $x_j = 0$ is $\ln\left(\frac{1}{1-P_j}\right) + E_j$ where E_j is exponential with rate 1.

$$\sum_{j=\ell+1}^{L} s_j = \ln(2) \cdot \underbrace{H_e^{[\ell+1:L]}(M,\pi,x)}_{} + \sum_{j=\ell+1}^{L} E_j$$

$$\geq H_e(M,\pi,x) - O(\lambda)$$

$$\geq \Omega(\lambda\sqrt{L}) \quad 2\lambda\sqrt{L-\ell}$$

$$\Pr\left[\sum_{j=\ell+1}^{L} s_j \leq (L-\ell) + \lambda\sqrt{L-\ell}\right]$$

$$= (L-\ell) - \lambda\sqrt{L-\ell} + 2\lambda\sqrt{L-\ell}\Big]$$

$$\leq \Pr\left[\sum_{j=\ell+1}^{L} E_j \leq (L-\ell) - \lambda\sqrt{L-\ell}\right]$$

$$\leq e^{-\lambda^2/2}$$

This comes from exponential r.v. tail bounds

$$\geq (L-\ell) - \lambda\sqrt{L-\ell}$$

$\Rightarrow$ decoding detection w.p. $\geq 1 - e^{-\lambda^2/2}$