Detecting Machine-Generated Texts by Multi-Population Aware Optimization for Maximum Mean Discrepancy

Shuhai Zhang[15*], Yiliao Song[2*], Jiahao Yang[1], Yuanqing Li[51†], Bo Han[4†], Mingkui Tan[13†]

South China University of Technology[1] The University of Adelaide[2]

Key Laboratory of Big Data and Intelligent Robot, Ministry of Education[3]
Department of Computer Science, Hong Kong Baptist University[4] Pazhou Laboratory[5]

April 18, 2024

Outline

Outline

Large language models (LLMs) have exhibited remarkable performance in generating human-like texts but may carry **critical risks**, *e.g.*, plagiarism issues, and hallucination issues.

Human or AI?

Background

It is challenging to distinguish machine-generated texts from human-written texts since **the distribution discrepancy between them is often very subtle** due to the advancement of LLMs.

Detector	HumanRec	MachineRec	AvgRec
GPT-3.5	96.98%	12.03%	54.51%
Human	61.02%	47.98%	54.50%

Li Y, Li Q, Cui L, et al. Deepfake text detection in the wild. arXiv, 2023.

Outline

Preliminaries of Maximum Mean Discrepancy

- *Maximum mean discrepancy* (MMD) aims to measure the distance between two distributions:

$$\begin{aligned}\mathrm{MMD}\left(\mathbb{P},\mathbb{Q};\mathcal{H}_k\right)&=\sup_{f\in\mathcal{F},\|f\|_{\mathcal{H}_k}\le 1}\left|\mathbb{E}[f(X)]-\mathbb{E}[f(Y)]\right|\\&=\sqrt{\mathbb{E}\left[k\left(X,X'\right)+k\left(Y,Y'\right)-2k(X,Y)\right]}.\end{aligned}$$

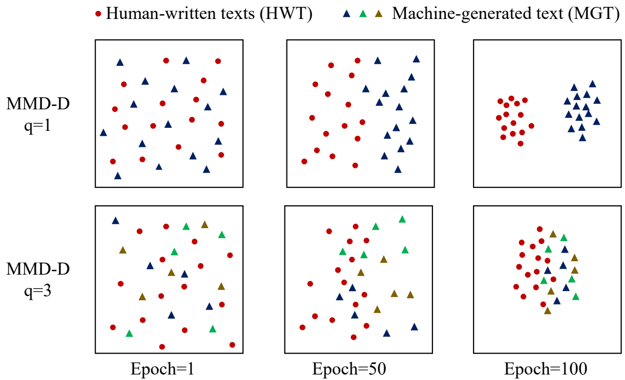
- Intuitively, we could view $k(X,X')$ or $k(Y,Y')$ as an *intra-class* distance and $k(X,Y)$ as an *inter-class* distance.
- Objective function for kernel-based MMD:

$$J(\mathbb{P},\mathbb{Q};k_\omega)=\mathrm{MMD}^2(\mathbb{P},\mathbb{Q};k_\omega)/\sigma_{\mathfrak{H}_1}(\mathbb{P},\mathbb{Q};k_\omega),$$

where $\sigma^2_{\mathfrak{H}_1}:=4\left(\mathbb{E}\left[H_{ij}H_{i\ell}\right]-\mathbb{E}\left[H_{ij}\right]^2\right)$ and
$H_{ij}:=k(\mathbf{x}_i,\mathbf{x}_j)-k(\mathbf{x}_i,\mathbf{y}_j)-k(\mathbf{y}_i,\mathbf{x}_j)+k(\mathbf{y}_i,\mathbf{y}_j)$.

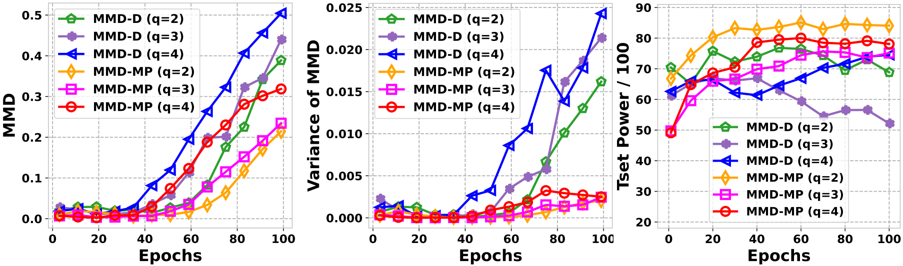
- For text detection, we define $\mathbb{P}$ and $\mathbb{Q}$ as the distribution of human-written texts (HWT) and machine-generated text (MGT), respectively.

High Variance Issue of MMD in Multiple Populations



- For MMD-D optimization, it tends to **separately aggregate** HWTs and **all possible** MGTs, such as decreasing the intra-class distance, and simultaneously push them away from each other like increasing inter-class distance.
- When the machine-generated texts population $S_{\mathbb{Q}}^{tr}$ **comprises different populations**, this optimization presents challenges due to **significant high variance issue of MMD**.

High Variance Issue of MMD in Multiple Populations



- During the optimization, as the number of $S_{\mathbb{Q}}^{tr}$ populations (*i.e.*, q) increases, kennel-based MMD (MMD-D) shows **an increase in MMD**, accompanied by **a sharp rise in variance**, resulting in **unstable test power** when testing.
- We propose a novel *multi-population* aware optimization method for MMD called MMD-MP, which exhibits **minimal variance of MMD values**, leading to **higher and more stable test power** when testing.

Outline

MMD-MP for Text Detection

Intuition for MMD-MP: During the training, we **do not consider optimizing the intra-class distance of machine-generated text samples** in $S_{\mathbb{Q}}^{tr}$.

- MMD-MP maximizes the following objective with a proxy MPP:

$$J(\mathbb{P},\mathbb{Q};k_\omega)=\mathrm{MPP}(\mathbb{P},\mathbb{Q};k_\omega)/\sigma_{\mathfrak{H}_1^*}(\mathbb{P},\mathbb{Q};k_\omega),$$

$$\mathrm{MPP}\left(\mathbb{P},\mathbb{Q};\mathcal{H}_k\right):=\mathbb{E}\left[k_\omega\left(X,X'\right)-2k_\omega(X,Y)\right].$$

- Empirically, we can estimate it by:

$$\hat{J}(S_{\mathbb{P}},S_{\mathbb{Q}};k_\omega)=\frac{\widehat{\mathrm{MPP}}_u(S_{\mathbb{P}},S_{\mathbb{Q}};k_\omega)}{\sqrt{\hat{\sigma}^2_{\mathfrak{H}_1^*}(S_{\mathbb{P}},S_{\mathbb{Q}};k_\omega)+\lambda}},$$

$$\widehat{\mathrm{MPP}}_u(S_{\mathbb{P}},S_{\mathbb{Q}};k_\omega)=\frac{\sum_{i\neq j}H^*_{ij}}{n(n-1)},\ \hat{\sigma}^2_{\mathfrak{H}_1^*}:=\frac{4}{n^3}\sum_{i=1}^n\left(\sum_{j=1}^n H^*_{ij}\right)^2-\frac{4}{n^4}\left(\sum_{i=1}^n\sum_{j=1}^n H^*_{ij}\right)^2,$$

$$\text{where } H^*_{ij}:=k_\omega(\mathbf{x}_i,\mathbf{x}_j)-k_\omega(\mathbf{x}_i,\mathbf{y}_j)-k_\omega(\mathbf{y}_i,\mathbf{x}_j).$$

Testing with MMD-MP for Text Detection

- Algorithm for paragraph-based detection, which can be considered as a two-sample test:

Algorithm 1 Testing with MMD-MP for 2ST

Input: Testing texts $S_{\mathbb{P}}^{te}$, $S_{\mathbb{Q}}^{te}$, $\hat{f}$, k_ω;

$est \leftarrow \widehat{\mathrm{MMD}}_u^2(S_{\mathbb{P}}^{te}, S_{\mathbb{Q}}^{te}; k_\omega)$;

for $i = 1, 2, \ldots, n_{perm}$ **do**

  Shuffle $S_{\mathbb{P}}^{te} \cup S_{\mathbb{Q}}^{te}$ into S_X and S_Y;

  $perm_i \leftarrow \widehat{\mathrm{MMD}}_u^2(S_X, S_Y; k_\omega)$;

end for

Output: p-value $\frac{1}{n_{perm}} \sum_{i=1}^{n_{perm}} \mathbf{1}(perm_i \geq est)$

Testing with MMD-MP for Text Detection

- Algorithm for sentence-based detection, which can be considered as a single-instance detection task:

Algorithm 2 Testing with MMD-MP for 2ST

Input: Referenced HWT $S_{\mathbb{P}}^{re}$, testing texts $S_{\mathbb{P}}^{te}$, $S_{\mathbb{Q}}^{te}$, $\hat{f}$, k_ω;
for $\mathbf{x}_i, \mathbf{y}_j$ in $S_{\mathbb{P}}^{te}, S_{\mathbb{Q}}^{te}$ **do**
  $P_i \leftarrow \widehat{\mathrm{MMD}}_b^2(S_{\mathbb{P}}^{re}, \{\mathbf{x}_i\}; k_\omega)$;
  $Q_j \leftarrow \widehat{\mathrm{MMD}}_b^2(S_{\mathbb{P}}^{re}, \{\mathbf{y}_j\}; k_\omega)$;
end for
Output: AUROC with two sets $\{P_i\}, \{Q_j\}$

Outline

Comparisons on Paragraph-based Detection

Our MMD-MP **achieves higher test power** than state-of-the-art text detection methods on HC3, given $3,100$ processed paragraphs in training data.

Method	ChatGPT	GPT3-S	Neo-S	ChatGPT Neo-S	ChatGPT GPT3-S
C2ST-S	$62.83_{\pm 0.90}$	$43.64_{\pm 5.92}$	$30.68_{\pm 2.37}$	$34.62_{\pm 2.73}$	$46.66_{\pm 2.95}$
C2ST-L	$89.82_{\pm 1.02}$	$75.74_{\pm 4.90}$	$60.97_{\pm 1.87}$	$68.50_{\pm 1.81}$	$78.22_{\pm 3.12}$
MMD-O	$26.43_{\pm 1.40}$	$21.17_{\pm 3.12}$	$19.83_{\pm 2.81}$	$25.23_{\pm 0.47}$	$25.18_{\pm 1.41}$
MMD-D	$91.76_{\pm 1.58}$	$86.98_{\pm 2.53}$	$75.45_{\pm 4.96}$	$86.44_{\pm 1.07}$	$91.46_{\pm 0.47}$
MMD-MP (Ours)	$\mathbf{93.21}_{\pm 1.35}$	$\mathbf{89.36}_{\pm 2.91}$	$\mathbf{79.68}_{\pm 2.42}$	$\mathbf{89.63}_{\pm 1.94}$	$\mathbf{91.96}_{\pm 0.62}$

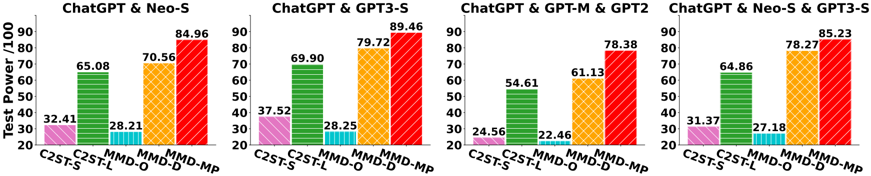
Comparisons on Sentence-based Detection

Our MMD-MP **outperforms state-of-the-art** methods for sentence-based detection on HC3 given $3,100$ processed paragraphs in training data.

Method	ChatGPT	GPT3-S	Neo-S	ChatGPT Neo-S	ChatGPT GPT3-S
Likelihood	$89.82_{\pm 0.03}$	$60.56_{\pm 1.32}$	$61.18_{\pm 1.25}$	$75.81_{\pm 0.51}$	$75.05_{\pm 0.25}$
Rank	$73.20_{\pm 1.49}$	$71.96_{\pm 1.01}$	$72.09_{\pm 0.51}$	$72.74_{\pm 0.74}$	$72.34_{\pm 1.38}$
Log-Rank	$89.58_{\pm 0.07}$	$63.78_{\pm 1.29}$	$64.92_{\pm 1.04}$	$77.57_{\pm 0.55}$	$76.47_{\pm 0.12}$
Entropy	$31.53_{\pm 0.90}$	$54.34_{\pm 1.33}$	$56.19_{\pm 0.33}$	$44.08_{\pm 0.24}$	$42.08_{\pm 2.01}$
DetectGPT-d	$77.92_{\pm 0.74}$	$53.41_{\pm 0.41}$	$52.07_{\pm 0.38}$	$66.01_{\pm 0.29}$	$65.70_{\pm 1.14}$
DetectGPT-z	$81.07_{\pm 0.77}$	$53.45_{\pm 0.53}$	$52.28_{\pm 0.31}$	$67.54_{\pm 0.19}$	$67.32_{\pm 1.02}$
OpenAI-D	$78.57_{\pm 1.55}$	$84.05_{\pm 0.71}$	$84.86_{\pm 0.87}$	$81.20_{\pm 0.95}$	$80.68_{\pm 1.64}$
ChatGPT-D	$95.64_{\pm 0.13}$	$61.89_{\pm 1.04}$	$54.45_{\pm 0.10}$	$75.47_{\pm 0.63}$	$78.95_{\pm 1.00}$
CE-Classifier	$96.19_{\pm 0.17}$	$92.44_{\pm 0.63}$	$88.88_{\pm 0.19}$	$90.93_{\pm 0.72}$	$92.97_{\pm 0.28}$
MMD-O	$56.34_{\pm 0.66}$	$59.90_{\pm 0.87}$	$63.19_{\pm 0.76}$	$60.46_{\pm 1.28}$	$57.79_{\pm 1.25}$
MMD-D	$95.83_{\pm 0.37}$	$94.86_{\pm 0.48}$	$91.12_{\pm 0.38}$	$91.39_{\pm 0.86}$	$93.49_{\pm 0.46}$
MMD-MP (Ours)	$\mathbf{96.20}_{\pm 0.28}$	$\mathbf{95.08}_{\pm 0.32}$	$\mathbf{92.04}_{\pm 0.58}$	$\mathbf{92.48}_{\pm 0.37}$	$\mathbf{94.61}_{\pm 0.22}$

Comparisons on Unbalanced Training Data

Our MMD-MP exhibits **significantly superior performance** compared with other methods, *e.g.*, surpassing the test power of $6.96\% \sim 14.40\% \uparrow$ than MMD-D, highlighting its stability under unbalanced training data scenarios.



Outline

Conclusions

- We delve into the optimization mechanism of MMD and reveal that **high variance of the MMD** when handling training data from **multiple different populations** can result in an **unstable discrepancy estimation** for MGT detection.
- We propose a novel multi-population aware optimization method for training kernel-based MMD (called MMD-MP), which can **alleviate the poor optimization of MMD-D** and **improve the stability of discrepancy measures**.
- Relying on MMD-MP, we develop two methods for **paragraph-based** and **sentence-based detection**, respectively. Extensive experiments across numerous LLMs, including ChatGPT, GPT2 series, GPT3 series, GPT-Neo series, demonstrate **superior detection performance**.

Thank you for your attention. Scan for more details.

Code

Paper