تمرین تئوری سری چهارم

امیرمحمد عزتی – ۴۰۲۲۱۲۲۶۹

سوال یک:

۱. روش NES بر اساس روش حداقل مربعات(Least Squares) عمل می‌کند. این روش به طور ریاضی اثبات شده که برای برآورد گرادیان در شرایط خاصی (مثل وجود نویز گوسی) بهینه‌ترین روش است. دلایل اصلی کم بودن واریانس این روش این‌طور اثبات شده است که:

NES و روش حداقل مربعات نتایج یکسانی دارند (این در قضیه ۱ اثبات شده است). همچنین روش حداقل مربعات اثبات شده بهینه با نمونه‌های محدود(finite-sample efficient) بوده و minimum-variance unbiased(MVU) است یعنی کمترین واریانس را در میان تمامی برآوردگرهای بدون اریب دارد (قضایای ۲ و ۳). در نتیجه، NES اطلاعات گرادیان را با استفاده از تعداد محدودی پرس‌وجو به بهترین شکل ممکن استخراج می‌کند.

۲. مقاله دو نوع اصلی دانش پیشین درباره گرادیان را معرفی می‌کند:

دانش پیشین(Prior) وابسته به زمان:

گرادیان‌های متوالی در روش‌های iterative مانند (PGD) معمولاً همبستگی بالایی با یکدیگر دارند. این همبستگی به‌طور تجربی از طریق شباهت کسینوسی مشاهده شده است که معمولاً حدود ۰.۹ باقی می‌ماند. این ویژگی یک ساختار قابل پیش‌بینی را فراهم می‌کند که می‌توان از آن بهره برد.

دانش پیشین(Prior) وابسته به داده:

در داده‌های تصویری، گرادیان‌ها معمولاً محلی بودن مکانی (Spatial Locality) نشان می‌دهند. پیکسل‌های نزدیک به هم اغلب مقادیر گرادیان مشابهی دارند. با این ویژگی ساختاری میتوانیم ابعاد مسئله را کاهش دهیم (مانند استفاده از روش "tiled" یا عملگر Avg Pooling) و در عین حال جهت کلی گرادیان حفظ کنیم.

هر دو نوع دانش پیشین باعث می‌شوند که تخمین گرادیان با تعداد کمتری پرس‌وجو کارآمدتر شود.

۳. تعریف کلی مسئله بندیت: مسئله بندیت یک چارچوب تصمیم‌گیری تکرارشونده است که در آن عامل(agent) باید در چندین مرحله اقداماتی را برای کاهش زیان تجمعی یا افزایش پاداش انتخاب کند.

در هر مرحله سه کار انجام انجام میشود به این صورت که ابتدا عامل(agent) با توجه به دانش خود یک اقدام عمل انجام می‌دهد. سپس زیان (یا پاداش) ناشی از آن عمل را مشاهده می‌کند و در نهایت دانش خود را برای بهبود عمل‌های آینده به‌روزرسانی می‌کند.

در روش پیشنهادی:

دانش پیشین: بردار نهان v_t که شامل دانش پیشین وابسته به زمان و داده است، دانش فعلی agent درباره گرادیان را کدگذاری می‌کند.

خروجی: تخمین گرادیان g_t است که با نگاشت v_t به فضای معتبر perturbationها بدست می‌آید.

تابع هزینه: l_t قرینه ضرب داخلی گرادیان تخمینی و گرادیان حقیقی، یعنی $-\left\langle \frac{g}{\|g\|}, \nabla L(x,y) \right\rangle$، که با استفاده از روش تفاضلات محدود تقریب زده می‌شود، این تابع کیفیت هم‌ترازی گرادیان تخمینی با گرادیان حقیقی را اندازه‌گیری می‌کند.

٤. الگوریتم حمله L_2 به‌صورت تکرارشونده ورودی خصمانه را با استفاده از تخمین گرادیان مبتنی بر بندیت اصلاح می‌کند.

Algorithm 2 Single-query spherical estimate of $\nabla_v \langle \nabla L(x,y), v \rangle$

procedure GRAD-EST(x, y, v)
  $u \leftarrow \mathcal{N}(0, \frac{1}{d} I)$ // Query vector
  $\{q_1, q_2\} \leftarrow \{v + \delta \boldsymbol{u}, v - \delta \boldsymbol{u}\}$ // Antithetic samples
  $\ell_t(q_1) = -\langle \nabla L(x,y), q_1 \rangle \approx \frac{L(x,y) - L(x+\epsilon \cdot q_1, y)}{\epsilon}$ // Gradient estimation loss at q_1
  $\ell_t(q_2) = -\langle \nabla L(x,y), q_2 \rangle \approx \frac{L(x,y) - L(x+\epsilon \cdot q_2, y)}{\epsilon}$ // Gradient estimation loss at q_2
  $\boldsymbol{\Delta} \leftarrow \frac{\ell_t(q_1) - \ell_t(q_2)}{\delta} \boldsymbol{u} = \frac{L(x+\epsilon q_2, y) - L(x+\epsilon q_1, y)}{\delta \epsilon} \boldsymbol{u}$
  // Note that due to cancellations we can actually evaluate $\boldsymbol{\Delta}$ with only two queries to L
  return $\boldsymbol{\Delta}$

مراحل به شرح زیر است:

مراحل اصلی:

١. مقداردهی اولیه: با تصویر اولیه x_0 و بردار نهان اولیه v_0 که دانش پیشین را کدگذاری می‌کند، شروع میکنیم.

٢. تخمین گرادیان (از طریق چارچوب بندیت):

در هر تکرار t، دو نسخه آشفته‌شده از v_t تولید میکنیم، $v_t + \delta u$ و $v_t - \delta u$، که u یک بردار گوسی تصادفی است. از این نسخه‌های آشفته برای تخمین مشتق جهت‌دار Δ_t با استفاده از روش تفاضلات محدود استفاده میکنیم:

$$\Delta_t = \frac{L(x + \epsilon(v_t - \delta u), y) - L(x + \epsilon(v_t + \delta u), y)}{\delta \epsilon} \cdot u$$

٣. به‌روزرسانی بردار نهان(latent):

v_t را با یک گام گرادیان صعودی تنظیم میکنیم:

$$v_t = v_{t-1} + \eta \Delta_{t-1}$$

٤. تولید اختلال:

از بردار به‌روز شده برای پیشنهاد گرادیان g_t استفاده کنید و آن را به تصویر اعمال میکنیم:

$$x_{t+1} = x_t + h \frac{g_t}{\|g\|_2}$$

مطمئن شوید که اختلال در محدوده نُرم L_2 باقی می‌ماند.

٥. تکرار تا موفقیت:

این فرآیند را ادامه می‌دهیم تا زمانی که طبقه‌بندی‌کننده ورودی آشفته x_t را اشتباه طبقه‌بندی کند.

سوال دو:

۱. بزرگی گرادیان ورودی(Size of Input Gradients): انتقال‌پذیری به میزان آسیب‌پذیری ذاتی مدل هدف در برابر نمونه‌های خصمانه وابسته است. مدل‌هایی با گرادیان بزرگ‌تر یا داشتن منظم‌سازی(regularization) ضعیف‌تر آسیب‌پذیری بیشتری دارند.

ترازبندی گرادیان‌ها(Gradient Alignment): شباهت (ترازبندی کسینوسی) بین گرادیان‌های مدل جانشین و مدل هدف تأثیر مستقیمی بر انتقال‌پذیری دارد. ترازبندی بهتر تضمین می‌کند که نمونه‌های خصمانه طراحی‌شده برای مدل جانشین در مدل هدف نیز مؤثر باشند.

تغییرپذیری تابع هزینه(Variability of the Loss Landscape): تنوع و تغییرپذیری کم تابع هزینه در مدل جانشین کمک می‌کند تا نمونه‌های خصمانه بهتر تعمیم یابند و احتمال موفقیت در انتقال افزایش یابد.

۲. خیر. مدل هدف ممکن است نمونه خصمانه را در همان کلاس مدل جانشین طبقه‌بندی نکند. این موضوع دلایل مختلفی میتواند داشته باشد.

عدم تطابق گرادیان: اگر گرادیان‌های مدل جانشین و مدل هدف به خوبی تراز نباشند، نمونه ممکن است به درستی منتقل نشود.

تفاوت‌های مدل: تفاوت در معماری، پیچیدگی و تنظیم منظم‌سازی بین مدل‌ها باعث ایجاد مرزهای تصمیم‌گیری متفاوت می‌شود.

سطح اطمینان: نمونه‌های خصمانه‌ای که confidence بالاتری در مدل جانشین دارند، ممکن است بهتر منتقل شوند، اما همچنان ممکن است به دلیل تغییر مرزها در دو مدل، با کلاس هدف یکسان طبقه‌بندی نشوند.

سوال سه:

۱. مقاله مدل‌های تهدید اصلی زیر را برای مدل‌های مولد عمیق (DGMs) معرفی می‌کند:

A) اختلال در مدل: ایجاد اختلال در فرآیند تولید مدل برای تولید خروجی‌هایی با کیفیت پایین یا نمونه‌های غیرمنتظره.

حملات Poisoning: وارد کردن داده‌های مخرب در فاز آموزش برای تضعیف پارامترهای مدل یا افزودن backdoor.

حملات Evasion: طراحی ورودی‌های خصمانه در فاز آزمایش برای تولید خروجی‌های نامطلوب.

B) سرقت اطلاعات محرمانه: نفوذ به حریم خصوصی با استخراج داده‌های حساس یا کپی‌برداری از مدل.

حملات استنتاج عضویت(Membership Inference Attacks): تعیین اینکه آیا یک داده خاص بخشی از مجموعه داده‌های آموزشی بوده است یا خیر.

حملات استنتاج ویژگی(Attribute Inference Attacks): استنتاج ویژگی‌های خصوصی داده‌ها بر اساس ویژگی‌های عمومی در دسترس.

حملات استخراج مدل(Model Extraction Attacks): کپی‌برداری از عملکرد مدل یا تقریب توزیع داده‌های آموزشی آن.

همچنین دانش پیشین متخاصم می‌تواند در این روش‌های حمله متفاوت باشد. ممکن است به داده‌های آموزش یا الگوریتم آموزش، یا به وزن و پارامترهای مدل، یا به latent code، یا صرفا به خروجی generate شده مدل‌ها دسترسی داشته باشد.

۲. روش‌های دفاعی زیر را برای کاهش این حملات در مقاله پیشنهاد شده است:

دفاع از اجزای مدل

- Weight Normalization: وزن‌ها را دوباره پارامتری‌سازی می‌کند تا طول و جهت از هم جدا شوند و تعمیم‌دهی را بهبود می‌بخشد، اما ممکن است در طول آموزش باعث ناپایداری شود اگر Generator یا Discriminator غالب شوند.
- Dropout: به‌طور تصادفی نورون‌ها را در طول آموزش غیرفعال می‌کند تا از Overfitting جلوگیری کند و تعمیم‌دهی را بهبود بخشد. با این حال، روند آموزش را کند کرده و تصاویر تار تولید می‌کند، که به ایپاک بیشتری برای دستیابی به نتایج معقول نیاز دارد.
- DPSGD (Differentially Private Stochastic Gradient Descent): در طول آموزش به گرادیان‌ها نویز اضافه می‌کند تا از حملات حریم خصوصی مانند (Membership Inference) محافظت کند. مؤثر است اما هزینه‌های محاسباتی را افزایش داده و کیفیت نمونه‌ها را کاهش می‌دهد.
- Smooth VAEs:
  - Double Backpropagation: با اضافه کردن تنظیم گرادیان، VAEs را در برابر تغییرات ورودی یا latent مقاوم‌تر می‌کند.
  - Disentangled Representations: ابعاد latent را مستقل نگه می‌دارد و حساسیت به عوامل غیرمرتبط را کاهش می‌دهد و استحکام را بهبود می‌بخشد.
- Fine-Pruning: ترکیبی از Pruning (حذف نورون‌های غیرفعال) و Fine-Tuning (آموزش مجدد بر روی داده‌های clean) است تا از حملات Poisoning و Backdoor جلوگیری کند. اما این روش، Utility مدل را کاهش داده و هزینه‌های محاسباتی را افزایش می‌دهد.
- Change Model Architecture:
  - PrivGAN: چندین جفت Generator-Discriminator را آموزش می‌دهد تا تقریب توزیع داده‌ها را مختل کند و یک Adversary داخلی برای مقابله با حملات Membership Inference دارد.

- RoCGAN: مسیرهای اضافی برای اعمال محدودیت بر خروجی اضافه می‌کند و از ورودی‌های خصمانه دفاع می‌کند.
- PATE-GAN: GANها را با تجمیع خصوصی Teacher Ensembles ترکیب می‌کند تا Differential Privacy را با استفاده از Discriminator های Teacher-Student تضمین کند.

- Digital Watermarking: شناسه‌هایی (Digital Watermarks) را در پارامترها یا خروجی‌های مدل جاسازی می‌کند تا مالکیت را تأیید کند، بدون اینکه از سرقت مدل را جلوگیری کند.

دفاع از خروجی‌های مدل

- Output Perturbation: به نمونه‌های تولیدشده نویز (مانند Gaussian Noise یا Adversarial Noise) اضافه می‌کند تا تقریب توزیع داده‌ها را مختل کند و از حملات حریم خصوصی جلوگیری کند. نویز ممکن است کیفیت را کاهش داده و توسط مهاجمان حذف شود.
- Activation Output Clustering: ورودی‌های غیرمعمول را با خوشه‌بندی خروجی‌های لایه‌های نهان تشخیص می‌دهد و احتمالاً از حملات Backdoor و Evasion دفاع می‌کند. اما این روش به حافظه زیادی نیاز دارد و برای DGMs کمتر مؤثر است.

دفاع از داده‌های آموزشی

- Expanding Training Set: داده‌های واقعی یا افزوده‌شده (مانند Flipping و Zooming) بیشتری را اضافه می‌کند تا تعمیم‌دهی را بهبود داده و از حملات Membership Inference جلوگیری کند.
- Input Perturbation:
  - Linear Interpolation: با ایجاد ورودی‌های جدید از طریق Interpolation بین نمونه‌های موجود، بازنمایی‌های latent را مختل می‌کند.
  - Semantic Interpolation: ویژگی‌های معنایی (مانند رنگ مو، عینک) را تغییر می‌دهد تا تنوع ایجاد کرده و در برابر حملات Model Extraction محافظت کند.

۳. انواع دانش‌هایی پیشینی که ممکن است مهاجم داشته باشد در مقاله عبارتند از:

داده‌ها و الگوریتم آموزش: دسترسی به مجموعه داده یا دانش درباره روند آموزش.

پارامترهای مدل: دسترسی به معماری مدل، وزن‌ها یا اجزایی مانند مولد، تفکیک‌کننده، کدگذار یا رمزگشا.

کد نهفته(latent code): دسترسی مستقیم یا دانش غیرمستقیم از توزیع latent.

داده‌های تولیدشده: دسترسی به نمونه‌های خروجی تولیدشده توسط مدل از طریق API ها یا اشتراک عمومی.

اطلاعات کمکی: هرگونه داده عمومی موجود که می‌تواند به حمله کمک کند، مانند مجموعه داده‌های مرتبط یا دانش درباره مدل‌های مشابه.