# 議題討論

組員:408530004 潘甫翰、408530007 謝滯瑩、408530028 楊宗軒、408530036 潘巧書、408530044 穆韋潔

## 期中前議題

## ● 題目







#### [議題 34] 面對對抗性攻擊時,模型可信性的探討

由408530004 潘甫翰發表於2023年 03月 26日(週日) 11:12

AI 系統可能會受到對抗攻擊的影響,攻擊者通常利用各種技術手段,如對抗樣本、對抗生成網絡等,對 AI 系統的輸入進行修改,使得 AI 系統產生不正 確或不合理的預測或決策。像是在圖像識別方面,攻擊者可以將一張正常的圖像添加一些微小的扭曲或干擾,從而使 AI 系統錯誤地識別圖像中的物體 或資訊。想問如何測試和驗證 AI 系統在面對對抗攻擊時的可信性?

永久鏈接 回覆



#### 回應:[議題 34] 面對對抗性攻擊時,模型可信性的探討

由409410024 陳品希發表於2023年 03月 29日(週三) 10:32

可以透過以不同方式攻擊AI系統,測試AI系統面對不同強度攻擊的反應;

也可使用對抗生成網路(Adversarial Generative Network,AGN),生成對抗樣本(adversarial examples),以測試和評估機器學習模型的強韌 性和可信度。

永久鏈接 顯示上層文章 回覆

[議題14] Pocket 演算法的 max iterations 應該設置多少才能在時間與最佳解取得平衡?

[議題15] 機器學習所帶來的隱私問題 ▶



[議題14] POCKET 演算法的 MAX ITERATIONS 應該設置多少才能在時間與最佳解取得平衡?由408530036 潘巧書發表於2023年 03月 23日(週四) 21:54

在作業一中有實作 Pocket 演算法,想知道 Pocket 演算法的 max iterations 大家都大概設定多少?

應該設置多少才能在時間與最佳解取得平衡呢?

歡迎大家提出各自的想法一起討論~



回應: [議題14] POCKET 演算法的 MAX ITERATIONS 應該設置多少才能在時間與最佳解取得平衡? 由611410036 陳振遠發表於2023年 03月 25日(週六) 18:26

觀察測試數據的模型效能,比較不同iterations值下的準確率及訓練時間,從中不斷優化

永久鏈接 顯示上層文章 回覆

## 回應議題



## 回應:[議題 34] 面對對抗性攻擊時,模型可信性的探討

由409410024 陳品希發表於2023年 03月 29日(週三) 10:32

可以透過以不同方式攻擊AI系統,測試AI系統面對不同強度攻擊的反應;

也可使用對抗生成網路(Adversarial Generative Network,AGN),生成對抗樣本(adversarial examples),以測試和評估機器學習模型的強韌 性和可信度。

永久鏈接 顯示上層文章 回覆



#### 回應: [議題 34] 面對對抗性攻擊時,模型可信性的探討

由408530004 潘甫翰發表於2023年 03月 30日(週四) 10:54

感謝您的回覆,我去查了資料,這兩種方法都是目前應用較廣泛的方法,並且確實可以提供有價值的測試結果,另外我也有查到,可解釋性 模型,也有助於提高模型的安全性和可靠性。

> 永久鏈接 顯示上層文章 編輯 刪除 回覆







#### 回應: [議題4]

由611410114 彭麟翔發表於2023年 03月 22日(週三) 22:22

儘管採用加密方式可以確保參數安全,但因廳大的計算量導致無法廣泛應用。因此有人提出差分隱私(Differential Privacy)的概念,透過產生隨機性的答案來遮蔽敏感資料。2016年Google提出一種差分隱私隨機梯度下降算法(SGD),將差分隱私(Differential Privacy)應用於神經網路的訓練上,能在適度的隱私預算下,在MNIST、CIFAR-10資料集上維持相對高的準確率,該方法目前已實作於Tensorflow\_privacy工具包。

永久鏈接 顯示上層文章 回覆



## 回應: [議題4]

由408530028 楊宗軒發表於2023年 03月 26日(週日) 13:39

同意,差分隱私提供一個隱私框架,確保組織能從這些資料中獲取資訊,但是又無法區分或是重新辨識出個人資訊,在數據分析、機器學習、社交網絡等領域得到了 廣泛的應用。

永久鏈接 顯示上層文章 回覆

# ● 回應其他組別發表的內容





# 組員:408530004 潘甫翰、408530028 楊宗軒、408530036 潘巧書、408530044 穆韋潔

## 期中後議題

## ● 題目

[議題373] AI對法律帶來的疑慮和挑戰 ◀ [議題415] 關於自駕車 [議題436] Transformer是否取代 CNN、RNN ▶ 以縮排方式呈現回應的貼文 設定 🗸 [議題373] AI對法律帶來的疑慮和挑戰 由408530044 穆韋潔發表於2023年 05月 29日(週一) 16:56 AI的快速發展,對法律帶來一些疑慮與挑戰, 例如: ● 隱私保護方面:Al需要大量的數據來訓練,這些數據可能包含個人隱私資料、醫療記錄、金融數據等敏感資訊。保護這些數據的隱私和安全成為一個重 要問題,因此需要相關的法律和監管措施 責任負擔方面:Al的自主性和自主決策能力使得確定法律責任變得複雜,當Al出現錯誤或造成損害時,決定誰應該負責並承擔法律責任成為一個挑戰,因 此法律需要明確界定AI和相關利益方之間的責任和義務。 除了上述舉例之外,可能還存在其他的法律相關問題, 歡迎各位同學提出看法共同討論~ 永久鏈接 回覆 OBJ OBJ [議題370]使用生成式AI和單純建立資料集訓練model的差異 ◀[議題353]機器學習應用於量化交易 [議題361]機器學習課堂差異▶ 以縮排方式呈現回應的貼文 設定 🗸 [議題370]使用生成式AI和單純建立資料集訓練MODEL的差異 由408530036 潘巧書發表於2023年 05月 28日(週日) 23:21 目前生成式AI正夯,但許多生成式AI的應用(例如:知識庫問答)看起來只是為了套上生成式AI而使用,實際上單純建立資料集訓練model可能就可以完成。 那什麼時候該使用生成式AI或是單純建立資料集訓練model就好呢?兩者有什麼差異呢? 永久鏈接 回覆 [議題371]CNN中FILTER如何學習



永久鏈接 回覆



## ● 回應議題



#### [議題373] AI對法律帶來的疑慮和挑戰

由408530044 穆韋潔發表於2023年 05月 29日(週一) 16:56

AI的快速發展,對法律帶來一些疑慮與挑戰,

#### 例如:

- 隱私保護方面:Al需要大量的數據來訓練,這些數據可能包含個人隱私資料、醫療記錄、金融數據等敏感資訊。保護這些數據的隱私和安全成為一個重要問題,因此需要相關的法律和監管措施。
- 責任負擔方面:Al的自主性和自主決策能力使得確定法律責任變得複雜,當Al出現錯誤或造成損害時,決定維應該負責並承擔法律責任成為一個挑戰,因此法律需要明確界定Al和相關利益方之間的責任和義務。

除了上述舉例之外,可能還存在其他的法律相關問題,

歡迎各位同學提出看法共同討論~

永久鏈接 回覆



#### 回應: [議題373] AI對法律帶來的疑慮和挑戰

由408530043 詹子賢發表於2023年 05月 31日(週三) 14:54

在科技日新月異情況下,AI將會應用在許多領域,在生活中有很多大小事都有AI參與時,我也認為會引發許多倫理、法律爭議。

就好比LEVEL-5的全自動自駕車,會自動收集外部環境資訊,來決定最佳路徑等,就可能涉及行人隱私問題。

若發生車禍事件,例如把前方拖板車畫面解釋成上交流道,因而加速導致碰撞,但責任歸屬會是AI還是駕駛者不應該將責任交屬於自駕車呢?

如今歐盟等國際團體也陸續推出AI責任指令草案,希望在科技進步的情況下,社會、法律、人民使用科技素質都能一同提升。

永久鏈接 顯示上層文章 回覆



#### 回應:[議題373] AI對法律帶來的疑慮和挑戰

由408530044 穆韋潔發表於2023年 05月 31日(週三) 16:28

#### 感謝同學的回覆!

隨著AI的發展,確實需要持續關注其帶來的倫理、法律和社會影響,

而這需要多方合作,包括科技公司、政府、學術界和公眾的參與,以建立合適的監管機制和法律框架,確保AI的應用符合人類價值觀和法律準則,並為人們帶來實際的結處。

永久鏈接 顯示上層文章 回覆



## 🦱 [議題370]使用生成式AI和單純建立資料集訓練MODEL的差異

由408530036 潘巧書發表於2023年 05月 28日(週日) 23:21

目前生成式AI正夯,但許多生成式AI的應用(例如:知識庫問答)看起來只是為了套上生成式AI而使用,實際上單純建立資料集訓練model可能就可以完成。

那什麼時候該使用生成式AI或是單純建立資料集訓練model就好呢?兩者有什麼差異呢?

永久鏈接 回覆



### 回應:[議題370]使用生成式AI和單純建立資料集訓練MODEL的差異

由409410062 洪偉誠發表於2023年 05月 31日(週三) 20:53

生成式AI和一般用資料集訓練的模型個人認為最大的差異在於學習資訊的方式不同,生成式AI在訓練中通常會使用大量的無標記資料來訓練,也就說是透過自行學習的方式來 從資料集中嘗試學出一些結構、特徵,而不像單純建立有標記的資料集來訓練,好讓模型可以更專注的去學習標記區塊的特徵、關聯,也正因為這種學習方式使生成式AI的訓 練結果通常更具有"創造力",因為模型甚至可能可以學習到人類無法輕易發現的相關性。

總之就我看來通常需要創造力的問題,像是產生圖片、文章,使用生成式AI或許就會是一個比較好的選擇,而相反的像是知識庫問答這類可能都有標準答案的問題,或許就不 是那麼適合。

永久鏈接 顯示上層文章 回覆



## 回應:[議題370]使用生成式AI和單純建立資料集訓練MODEL的差異

由408530036 潘巧書發表於2023年 05月 31日(週三) 21:01

原來如此,謝謝你的回答!

永久鏈接 顯示上層文章 回覆

#### 🦱 [議題371]CNN中FILTER如何學習

由408530028 楊宗軒發表於2023年 05月 29日(週一) 11:48

請問CNN中的filter是如何學習的?

永久鏈接 回覆

## 回應: [議題371]CNN中FILTER如何學習

由408530043 詹子賢發表於2023年 05月 31日(週三) 15:03

CNN的核心是卷積層,其中包含許多filter,每一個filter都是一個小型權重矩陣,形狀通常為方形,並透過滑動窗口方式,在圖像上進行運算、訓練。

在訓練過程中,CNN通過反向傳播法進行運算,步驟包括隨機初始化filter的權重值、進行向前傳播、計算loss function、進行反向傳播、使用梯度下降法更新filter權重值。透過反覆執行上述步驟,CNN能夠學習到最佳 filter權重值,從而更好題曲圖像特徵。

永久鏈接 顯示上層文章 回覆



回應: [議題371]CNN中FILTER如何學習 由☆ 江振國發表於2023年 06月 2日(週五) 00:06

CNN中的filter就是模型要學習的權重或模型參數,你可以參考投影片上說明,將filter的權重值對應到全連結的權重值,這樣就可以利用跟DNN一樣的反向傳播演算法學習了~

永久鏈接 顯示上層文章 回覆



#### ■ 回應: [議題371]CNN中FILTER如何學習

由408530028 楊宗軒發表於2023年 06月 2日(週五) 14:04

原來如此,謝謝老師和同學的回覆~

永久鏈接 顯示上層文章 回覆



#### 回應: [議題 399] 黑箱攻擊時模型的決定

由611410081 陳楷勛發表於2023年 06月 3日(週六) 16:22

回覆第一個問題,我有想到類似的技術稱為模型提取攻擊(Model Extraction Attacks),在這種攻擊中,攻擊者會使用從目標模型得到的輸入/輸出 的pair來訓練一個新的模型,這個新模型被訓練用於模擬目標模型的行為。而第二個問題的部份可以參考對抗樣本攻擊(Adversarial Attacks),這 種方法可以在對目標模型知識有限的情況下進行的攻擊,攻擊者會生成一些特殊的輸入資料,這些輸入資料在人類看來與正常輸入差不多,例如存 在人類無法辨識的雜訊,可以誘導機器學習模型做出錯誤的預測。

永久鏈接 顯示上層文章 回覆



### 回應: [議題 399] 黑箱攻擊時模型的決定

由408530004 潘甫翰發表於2023年 06月 5日(週一) 19:24

非常威謝您的回覆!您提到的模型提取攻擊和對抗樣本攻擊方法似乎能夠用於推測目標模型的行為或者欺騙它。這些方法提供了有價值的策 略,以便在黑箱攻擊中採取行動。

至於您的第二個問題,根據我所了解,黑箱攻擊通常不需要完全猜中模型的所有細節。即使僅對模型的架構有一定的了解,也可能能夠訓練 出具有強大惡意行為的樣本。當然,對於了解模型的細節越多,可能能夠更準確地進行攻擊,但在實際應用中,攻擊者可能只需掌握模型的 大致架構就能展開相應的攻擊。

永久鏈接 顯示上層文章 回覆

# 回應其他組別發表的內容



#### [議題354] 能跟AI變成朋友嗎?

由408415037 張庭瑜發表於2023年 05月 22日(週一) 17:10

著疫情對AI需求更劇烈的助力下,人工智慧的相關軟體層出不窮,無論是自動繪圖軟體、自動影片產生AI、動畫產生AI、目前最夯的ChatGPT以及微軟的 Copilot…等等。

許多人認為AI 還不成氣候,ChatGPT 有時候給的答案是錯的,或是每次問給的答案都不盡相同;AI繪圖軟體常常畫出很奇怪的怪物,圖片轉換成動畫也是 有點肢體障礙。

事實上,未來的AI會越來越精進已不可避免,或許有一天他也能夠學習情緒。

永久鏈接 回覆

回覆



#### 回應: [議題354] 能跟AI變成朋友嗎?

由408530044 穆韋潔器表於2023年 05月 29日(週一) 16:32

現今其實已存在具有情緒的機器人,例如客服機器人、陪伴機器人等,它們能夠分析和解讀人的語言、語調、表情和肢體語言等,從而理解人的情緒 狀態,並以適當的方式回應。

但目前的情感機器人仍處於發展初期,雖然能夠模擬一些情緒表達和回應,但與人類的情感體驗相比還存在差距。

永久鏈接 顯示上層文章



#### [議題368] 期末AI桌遊

由409410024 陳品希發表於2023年 05月 28日(週日) 01:57

也快到期末了,想問大家有玩過AI桌遊了嗎?在玩的過程中有遇到什麼問題、或有什麼致勝方法可以分享嗎?

永久鏈接 回覆



回應: [議題368] 期末AI桌遊由408530036 で 由408530036 潘巧書發表於2023年 05月 31日(週三) 16:36

玩起來感覺運氣成分占一部分,除此之外也可以對機器學習更了解~

永久鏈接 顯示上層文章 回覆



#### 回應: [議題368] 期末AI桌遊

由☆ 江振國發表於2023年 06月 2日(調五) 00:03

偷偷告訴你 (希望别人不會看到...XD),這個遊戲有兩大致勝策略: 第一、你可以全部都蓋最低分的AI模型,但需要拿到最低分的AI模型最後能額外加分的那一張高階功能牌, 致勝關鍵是快速結束比賽,別人在蓋大型複雜的功能牌5-6張的時候,你已經提前結束比賽,且有大量額外加分;第二是你可以盡量全蓋功能牌,只有少量的AI模型用來訓 練、測試,但需要拿到功能牌能額外加分的那張高階功能牌,這個分數是有機會打敗前一種的,因為加分更高。很怕的是玩家沒有目標,東蓋西蓋沒有重點-

永久鏈接 顯示上層文章 回覆



回應:[議題368] 期末AI桌遊 由409410024 陳品希發表於2023年 06月 2日(週五) 13:17

原來如此,希望我的組旨不會看到XD, 感謝老師回覆!

永久鏈接 顯示上層文章 回覆



### [議題350] PYTORCH & TENSORFLOW 之使用經驗分享

由408420037 侯貝霖發表於2023年 05月 21日(週日) 22:03

PyTorch和TensorFlow是當前最流行的深度學習框架,兩者的差異在學期初課堂上也有稍微提到。基本上PyTorch被認為在易用性方面較TensorFlow更友好,使初學者較輕易上手,同時 我也發現在一些深度學習的相關競賽中,滿多的參賽者都是使用PyTorch作為框架。

因此,我想請問大家在一開始接觸這個深度學習的領域時是先學習使用何種框架? 覺得兩者使用起來的優缺點或適應性有什麼不同呢?

歡迎大家分享自身經驗~謝謝

永久鏈接 回覆



#### ■ 回應: [議題350] PYTORCH & TENSORFLOW 之使用經驗分享

由408530028 楊宗軒發表於2023年 05月 29日(週一) 16:03

我習慣使用PyTorch,而兩者的比較在於TensorFlow提供較好的視覺化工具,使用者較易debug以及追蹤training process,但是在資料平行處理方面,PyTorch透過 Python原生的非同步執行,能有較好的效能。

永久鏈接 顯示上層文章 回覆



[議題391]不小心使用惡意訓練資料集的問題 由408410098 蔡嘉祥發表於2023年06月1日(週四)13:39

會有有心人士在公開的訓練資料集之中植入惡意訓練資料,想導致模型使用後產生缺陷。

如何判斷使用的訓練資料集是否此狀況?

或是訓練時可以發現?



## 回應: [議題 391] 不小心使用惡意訓練資料集的問題 由408410070 蘇展陞發表於2023年 06月 1日(週四) 14:08

我覺得可以試著讓資料可視化,例如繪製資料的散佈圖,應該有助於發現某些異常的資料點。 或是應盡量使用來自較可靠來源的資料集,比較能夠避免此情況。

**ネタ雑技 顕示上層文音 同層** 



## 回應: [機蹈 391] 不小心使用惡意訓練資料集的問題 由408410102 楊力行發表於2023年 06月 5日(週一) 16:34

谢谢同學的回覆,有時即使經過可視化,惡意資料還是不好辨認,但使用公認的、有可靠來源的資料及確實能避免使用到惡意資料的情形。

永久鏈接 顯示上層文章 回覆



## 回應: [機距 391] 不小心使用惡意訓練資料集的問題 由☆ 江振園發表於2023年 06月 1日(週四) 23:44

目前的公園訓練集在跟式公開訓練集的時候,其質都需要投稿到頂賴會議中,並提供完整的測試驗證,說明資料的韓訊比例、指談標記等,並用SOTA的方法驗證資料集的率確性,對領域技術發展的未來有幫助,才會被接受,被接受之後才會公開。因此,如果是從正當的管路下藏的公開資料集,比較 這個免疫地入息查資料的問題,



## 回應: [議題 391] 不小心使用惡意訓練資料集的問題 由408530004 潘甫翰發表於2023年 06月 2日(週五) 01:51

我觉得,在公開訓練集發布的時候,可能也可以設計相關的認證機制(模型)來評判說該樣本是否為惡意樣本,若大家都可以對此樣本進行測試並驗證,那當樣本被惡意修改,在眾多使用者的測試把關下,我們找出問題的可能性應該也會變大。