敏求智慧運算學院
Miin Wu School of Computing

透過低層次特徵學習增進對抗性防禦模型的穩健性

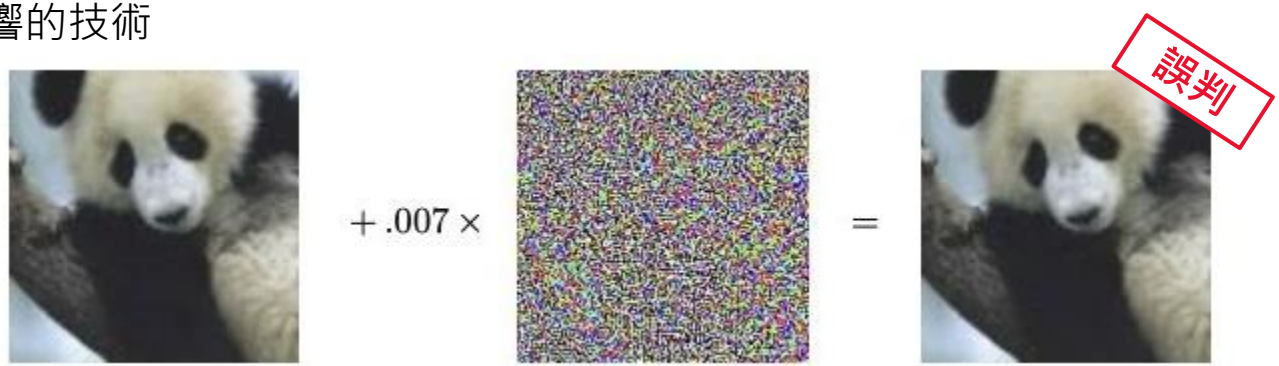
研究領域：電腦視覺中的對抗性防禦
作者：劉恩兆 成功大學統計學系四年級
指導教授：許志仲 成功大學統計學系助理教授

國立成功大學 第一屆 敏求智慧運算競賽
The Grand Challenge : Computing for the Future

大綱

前言
對抗性攻擊與對抗性防禦介紹

- 對抗性攻擊：指以演算法生成人們難以察覺，並且能干擾機器學習模型的對抗例（Adversarial Example，也稱為對抗性樣本），以此來引發機器學習模型誤判的一種攻擊手法
- 換句話說，對抗性攻擊的目標就是使模型誤判的同時，將圖片中的改動盡可能地控制在微小的範圍
- 對抗性防禦：是一種提高機器學習模型的穩健性，使其能夠抵抗或減少對抗性攻擊影響的技術



熊貓（原始乾淨圖像）
57.7%信心分數

Perturbation（擾動）

長臂猿（對抗性樣本）
99.3%信心分數

圖1、使用對抗性攻擊FGSM使深度學習模型誤判範例 [1]

研究動機
對抗性防禦是深度學習領域中關於安全性最重要的議題之一

- 現今深度神經網路受到廣泛應用，但僅有高準確率是不夠的，還需要克服其致命的缺點—容易被對抗性攻擊欺騙，導致錯誤預測或分類
- 因此，研究對抗性防禦方法可以幫助提升安全以及維持公平性
  - 人身安全：以自動駕駛為例，若是用來偵測道路及車輛的深度神經網路受到對抗性攻擊干擾，則可能造成自動駕駛車輛做出錯誤決策，導致車禍發生
  - 社會公平性：以人臉辨識應用為例，對抗性防禦能夠防止攻擊者在模型中加入偏見與漏洞，以避免種族、性別或是年齡等方面的歧視

停止號誌（乾淨圖像）　停止號誌（受攻擊）　速限號誌（誤判）

圖2、對抗性攻擊使自動駕駛辨識系統誤判範例 [2]

方法介紹
主流對抗性防禦原理介紹

- 現今主流的對抗性防禦方法大致可分為以下三類：
  1. 混淆梯度方法（Obfuscated Gradients Method）：藉混淆或是干擾梯度信息，使攻擊者無法通過計算模型的梯度獲得有效的對抗性樣本
  2. 對抗性訓練（Adversarial Training）：通過在模型訓練過程中預先加入對抗性樣本，經過反覆訓練以增加模型穩健性
  3. 基於輸入的轉換方法（Input Transformation Method）：在數據輸入模型之前先進行變換與處理，濾除或減弱攻擊對模型所帶來的影響
- 然而，這三類防禦方法皆各有其缺點與限制：

對抗性防禦方法	混淆梯度方法	對抗性訓練	基於輸入的轉換方法
缺點與限制	1. 一旦混淆梯度策略被攻擊者得知，即可以輕易通過近似或繞過攻擊算法的梯度攻擊成功 [3]	1. 需額外耗費大量運算資源 2. 無法防禦未經對抗性訓練的攻擊方法 3. 對乾淨圖像的辨識能力下降	1. 一旦輸入轉換的方法被得知，即可以視為神經網路第一層進行攻擊，導致防禦失敗 2. 對乾淨圖像的辨識能力下降

方法介紹
站在巨人的肩膀上—提出運算量低並防禦效果極佳的解決方案

- 欲解決現今防禦方法的痛點，可以從以下兩點著手: 1）降低運算量 2）提高防禦表現
- 而混淆梯度方法與基於輸入的轉換方法遇到最大的瓶頸為：攻擊者容易經過梯度計算使用BPDA（Backward Pass Differentiable Approximation）近似
- 因此，本文借鑒混淆梯度與輸入轉換的思想，提出使用簡單的傳統低層次特徵轉換方法：LBP（Local Binary Pattern）與LTP（Local Ternary Pattern），由這兩種高度不可微分的轉換函數，防止攻擊者經由梯度計算找到適合的擬合函數進行攻擊

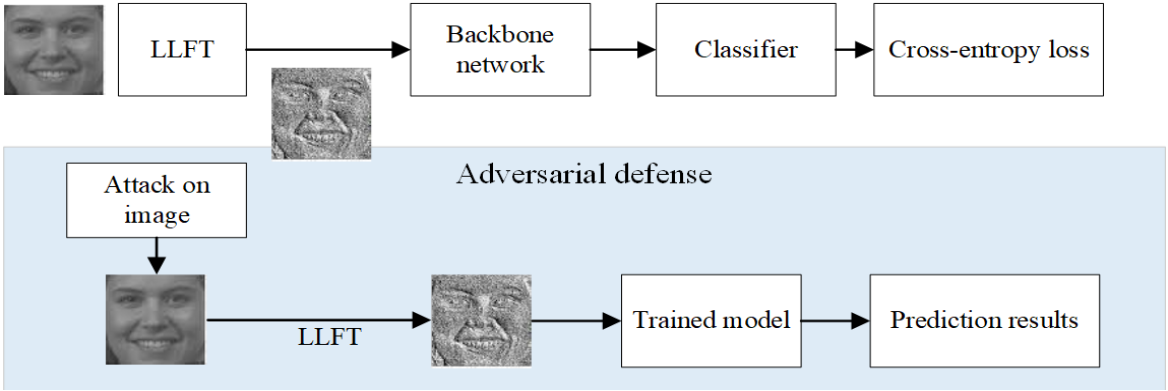


圖3、本文所提出的方法流程圖（使用低層次特徵轉換提升模型穩健性）

註：對抗性攻擊原理請見附錄（一）；BPDA原理請見附錄（二）

方法介紹
選擇低層次特徵轉換— LBP/LTP之原因

- 選擇條件：
  1. 輸入經過轉換後，對神經網路仍要具有良好的穩健性
  2. 不容易被BPDA（Backward Pass Differentiable Approximation）經過梯度計算近似成功
- 選擇低層次特徵之原因：
  1. 對抗性攻擊並非針對低層次特徵轉換而設計，因此真實的局部特徵與經過對抗性攻擊污染的局部特徵，在進行低層次特徵轉換後的特徵表示可能不同，使神經網路可以辨識出對抗性攻擊
  2. 低層次特徵使原始圖像和轉換後的圖像之間的領域差距更大，表示BPDA可能難以近似，從而使對於輸入轉換的攻擊無效
- 選擇LBP/LTP之原因：
  1. 具有非線性與離散性質：此兩種轉換為不平滑且不連續的轉換方式，無法使用常見的可微分方法（如：梯度下降）來進行最優化或近似，極大避免了受到BPDA近似攻擊
  2. 運算簡單、速度快：由後續實驗結果可知，在不明顯增加運算時間下也能增加模型穩健性
  3. 經過轉換後的信息損失不多：由實驗結果可知，經轉換輸入的熵直方圖與原始RGB的熵直方圖領域（Domain）非常相似，表示經轉換後的特徵與原始圖像之間的性能差距足夠小

註：LBP轉換介紹請見附錄（三）；LTP轉換介紹請見附錄（四）；熵直方圖請見附錄（五）

實驗結果
實驗設置

- 數據集：ImageNet
- 模型設置：
    - Epoch：90，每訓練完30個epochs會將學習率按照0.1的比例縮小
    - Optimizer：Adam，權重衰減為1×10^6
    - Batch size：256
    - 額外設置：因為本文的防禦方法的輸入與傳統深度神經網路不同，我們不使用任何預訓練權重來初始化本文的防禦方法的參數
- 硬體配置：i9-9900k Intel CPU和4個NVIDIA Tesla V100的個人電腦
- 訓練總花費時間：47小時
- LBP與LTP參數設置：R和P分別為3和8，LTP的閾值t則是從3到12進行抽樣
- 實驗共收集7種不同的白盒攻擊方法：PGD、FGSM [4]、AutoAttack、SPSA、TGSM、DeepFool和VMI-FGSM

註：7種白盒攻擊方法介紹請見附錄（六）

實驗結果
本文方法和其他先進防禦方法的性能（準確率）比較

Method	Clean	FGSM	PGD	AutoAttack	SPSA
ResNet-50	75.4	19.0/12.4/7.9	3.1/0.2/0.0	6.8/0.5/0.2	9.4/7.2/3.3
DIPDefend [5]	45.5	40.5/32.9/29.9	41.2/39.9/39.9	40.8/38.6/37.5	42.5/41.2/39.0
SAT	69.7	40.6/38.8/31.7	43.0/42.2/34.9	-	-
FD* [6]	74.8	48.5/43.7/35.9	49.6/45.5/34.9	44.7/41.5/37.5	50.6/46.3/39.8
Ours/LBP	67.2	44.9/37.8/30.4	48.5/47.6/47.0	67.1/67.1/67.0	60.4/58.3/55.4
Ours/LTP	65.4	59.9/52.0/36.7	60.2/55.1/47.5	64.5/64.4/64.6	60.3/58.4/56.2

表 1 、本文方法和其他先進防禦方法的性能（準確率）比較，其中FGSM、PGD、AutoAttack和SPSA攻擊的設置為 $\epsilon = 4/8/16$，PGD的 $\alpha = 2$ 和 $s = 50$，*表示由第三方實現

註：先進防禦方法介紹請見附錄（七）

實驗結果
本文方法在不同攻擊下的性能（準確率）比較

Param.	ϵ = 4/8/16			
Method	DeepFool	VMI-FGSM	TGSM	RFGSM
Ours/LBP	64.8/64.8/64.7	60.0/53.9/44.8	60.8/51.2/37.3	60.4/51.3/33.7
Result	Learned LTP / Perturbed LTP			
BPDA type	Naive	2-layer	4-layer	RRDB
Ours/LTP	-/53.5	58.9/47.1	59.1/46.2	57.2/43.5

表2、對本文方法在不同攻擊方法下的性能（準確率）進行比較，其中對抗性攻擊的三個結果是在不同的擾動水平 ϵ 下獲得的，而BPDA結果則是在使用PGD進行近似時在 $\epsilon = 8$、 $\alpha = 2$ 和 $s = 10$的情況下獲得的

結論

- 本文在不明顯增加推理時間的條件下，提出了一種快速推理感知的低層次特徵轉換防禦方法，結合輸入轉換和混淆梯度的技巧，使用LBP與LTP兩種低層次特徵轉換方式提高神經網路的穩健性，並證明此方法仍能獲得極優異的表現
- 實驗結果表1顯示，本文提出的LBP和LTP特徵轉換防禦方法在性能上明顯優於其他防禦方法，並且相較於進行對抗性訓練，本文所提出的防禦方法在訓練和推理階段具有較低的運算成本
- 實驗結果表2顯示，即使在肉眼可見的擾動程度，傳統的BPDA（Backward Pass Differentiable Approximation）仍無法很好地近似LTP轉換，導致攻擊無效
- 總結來說，此實驗證實了既不需要在線訓練，也不需要對抗性訓練，僅使用LBP與LTP這兩種低層次特徵轉換即可以顯著地提升模型的穩健性

具體貢獻
由簡單的傳統低層次特徵轉換達到極優異的防禦性能

- 關於創新貢獻，本文主要達成了以下兩點：
    1. 根據詳盡的調查，本文首次揭示了傳統低層次特徵在不明顯增加推理時間的情況下，有助於模型的穩健性
    2. 本文展示了所提出的方法在不同的對抗性攻擊下皆能取得最先進的性能，同時無需進行對抗性訓練造成運算資源浪費
- 關於社會影響，本文則達成了以下兩點：
    1. 解決對抗性訓練方法運算資源要求較高之痛點，成功以更低的運算資源獲得更佳的防禦表現
    2. 由本文提出的運算資源要求較低之方法，使對抗防禦方法更符合現實應用場景，讓更多社會大眾能夠一起參與、了解深度神經網路的安全領域

未來研究方向

- 引入其他低層次特徵（如：Dense SIFT（dense scale-invariant feature transform）），以驗證是否經過其他低層次特徵轉換，模型仍能具有穩健性
- 尋找其他適用於對抗性防禦的低層次特徵，其需滿足：
  - 較LBP/LTP轉換含有更豐富資訊的特徵轉換方式，使模型預測表現更佳
  - 較LBP/LTP運算更快速的特徵轉換方式

參考文獻

[1] Goodfellow, I.J., Shlens, J., Szegedy, C., “Explaining and Harnessing Adversarial Examples.” arXiv preprint arXiv：1412.6572（2014）

[2] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song ： Robust Physical-World Attacks on Deep Learning Models. arXiv preprint arXiv：1707.08945（2017）

[3] Anish Athalye, Nicholas Carlini, David Wagner ： Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. ICML（2018）

[4] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572 (2014)

[5] Dai, T., Feng, Y., Chen, B., Lu, J., Xia, S.T.: Deep Image Prior Based Defense Against Adversarial Examples. Pattern Recognition 122, 108249 (2022)

[6] Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature Denoising for Improving Adversarial Robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 501–509 (2019)

[7] Ahmed, N., Natarajan, T., Rao, K.R.: Discrete Cosine Transform. IEEE transactions on Computers 100(1), 90–93 (1974)

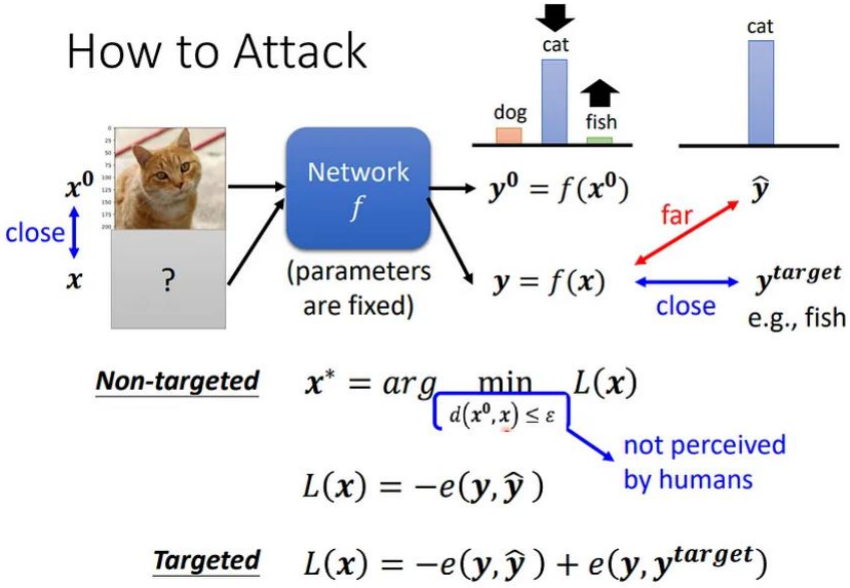
[8] Akhtar, N., Mian, A.: Threat of Adversarial Attacks on Deep Learning in Computer Vision: A survey. Ieee Access 6, 14410–14430 (2018)

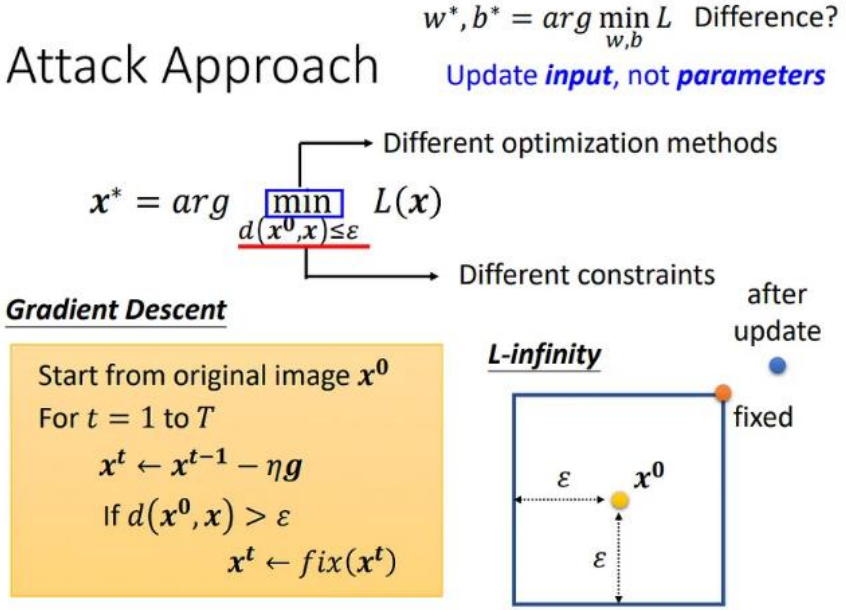
敏求智慧運算學院
Miin Wu School of Computing

Q & A

國立成功大學 第一屆 敏求智慧運算競賽
The Grand Challenge : Computing for the Future

附錄（一）
對抗性攻擊原理





Source from: 國立台灣大學李宏毅教授

附錄（二）
BPDA原理

- BPDA（Backward Pass Differentiable Approximation）是一種用於對抗性攻擊中的技術，用於處理梯度的不可微問題。在對抗性攻擊中，攻擊者通常需要計算模型對於對抗性樣本的梯度，以便生成有效的對抗性樣本。然而，有些模型可能包含了不可微分的操作，導致無法直接計算梯度。
- BPDA的基本思想是使用一種可微分的近似方法來替代不可微分的操作，以獲得有效的梯度估計。具體步驟如下：
    1. 正向傳遞：將原始輸入數據通過模型進行正向傳遞，獲得模型的預測結果。
    2. 反向傳遞近似：由於模型中包含不可微分的操作，無法直接計算梯度。在這裡，使用一種可微分的近似方法，將不可微分的操作替換為可微分的近似操作。這樣可以使得反向傳遞得以進行，並計算出近似的梯度。
    3. 梯度更新：使用近似的梯度信息來更新對抗性樣本，使其更接近讓模型產生誤判的數據點。
- BPDA的優勢在於通過使用近似方法解決了梯度不可微的問題，使得攻擊者能夠在包含不可微分操作的模型上進行有效的對抗性攻擊。然而，由於近似操作的使用，BPDA可能會引入一定的誤差，這可能影響攻擊的效果。

附錄（三）
LBP轉換介紹

- LBP（Local Binary Pattern）
  - 定義為：$LBP_{P,R} = \sum_{p=0}^{P-1} 2^p s(x_p - x),\ s(x) = 1\ if\ x \geq 0\ else\ 0$
    其中，P是以像素 x 為中心的周圍像素數量，由半徑R決定

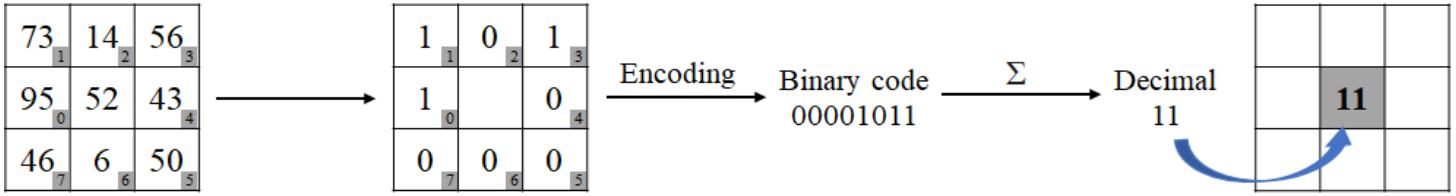
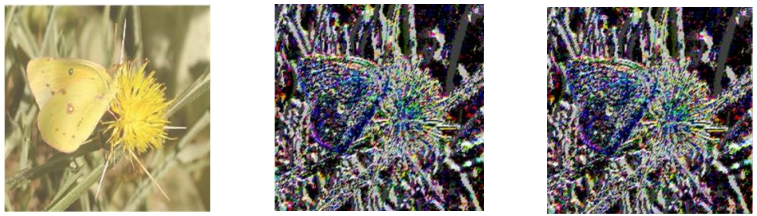


圖4、LBP運算子簡介圖



Clean　　LBP　　LTP

圖5、原始圖像與LBP、LTP轉換圖像比較

附錄（四）
LTP轉換介紹

- LTP（Local Ternary Pattern）
  - 定義為：$LTP_{P,R} = \sum_{p=0}^{P-1} 2^p T(x_p - x),\ T(x) = \begin{cases} 1, x \le t \\ 0, -t < x < t \\ -1, x < t \end{cases}$

    其中，P是以像素 x 為中心的周圍像素數量，而$\mathrm{T}(x)$是一個三元函數，根據$x_p - x$的值返還0、1或-1，由閾值t決定

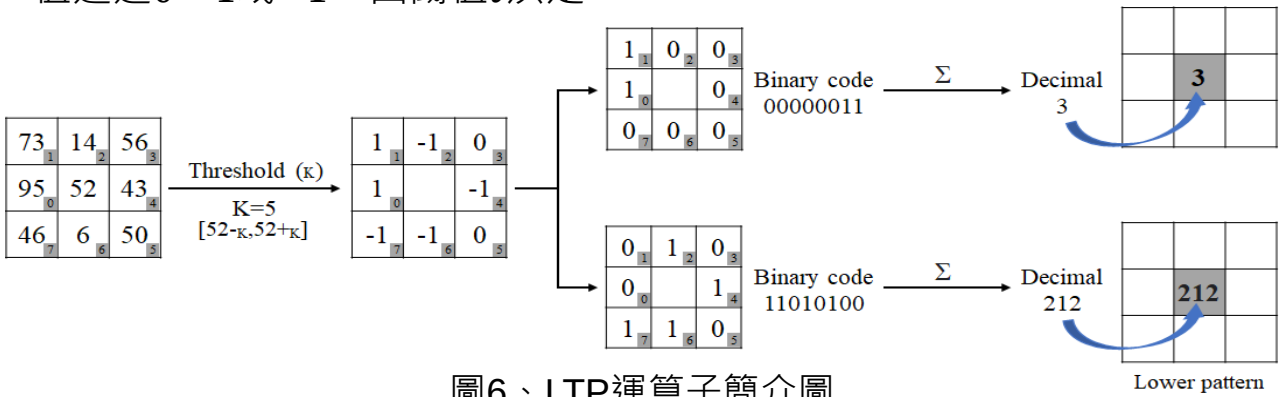


圖6、LTP運算子簡介圖

附錄（五）
對抗性攻擊實證指標介紹

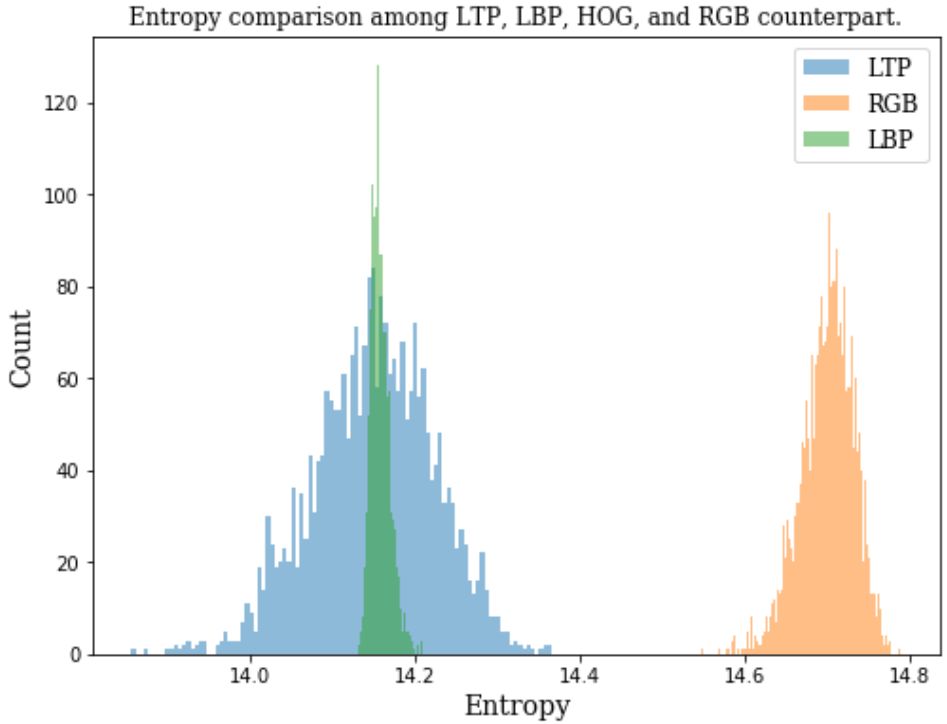


圖 6、RGB與經LBP/LTP轉換之熵直方圖

附錄（六）
對抗性攻擊實證指標介紹

- 一些典型的攻擊方法被廣泛用作實證指標，用於評估各種防禦方法的有效性：
  1. FGSM（Fast Gradient Sign Method）：通過計算損失函數的梯度來生成對抗性樣本，需要通過反向傳播獲得真實的梯度
  2. PGD（Project Gradient Descent）：是一種常見的迭代式對抗性攻擊方法，透過在每次迭代中計算模型的梯度並更新輸入樣本，產生多個對抗性樣本，直到滿足攻擊目標為止，被認為是利用局部一階信息的最強攻擊方法
  3. AutoAttack ：是一種結合了多種攻擊方法的對抗性攻擊方法，其中包括兩個新版本的PGD、C&W（Carlini and Wagner Attacks）和Square Attack等
  4. SPSA（Simultaneous Perturbation Stochastic Approximation）：是一種基於隨機梯度估計的對抗性攻擊方法，會重複進行攻擊迭代直到滿足最大攻擊迭代次數或達到特定的攻擊成功率
  5. TGSM（Targeted Gradient Sign Method）：在決定目標後計算對抗性樣本的梯度，並使用梯度的符號來決定對抗性樣本中的像素值改變的方向，並重複迭代改變像素值，以增加攻擊的效果
  6. DeepFool：初始化後計算模型在對抗性樣本上的梯度，並選取使模型改變的最短距離，後不斷迭代。其特點在於能夠生成極小的干擾，使得對抗性樣本在視覺上很難被人類察
  7. VMI-FGSM（Virtual Momentum-based Iterative Fast Gradient Sign Method）：為FGSM的延伸，會引入虛擬動量，將梯度的方向和大小結合起來，使攻擊更加穩定和高效

附錄（七）
對抗性防禦實證指標介紹

- 一些表現優異的對抗性防禦方法會被廣泛用作評比效能的驗證指標：
  1. DIPDefend（Deep Image Prior Defense）：其利用深度圖像先驗知識，通過對輸入圖像進行重建和修復來抵抗對抗攻擊，提高深度學習模型的韌性和抵抗能力
  2. SAT（Smooth Adversarial Training）：其為一種對抗性訓練方法的變體，與傳統的對抗性訓練不同，SAT在生成對抗性樣本時引入了平滑化的步驟，通過給對抗性樣本加入一定程度的隨機噪聲，使其在視覺上更加接近原始數據
  3. FD（Feature Denoising）：通過在模型的特徵層面引入噪聲或篩檢等處理，使攻擊者難以在特徵層面找到有效的攻擊方向