機器學習 hw10 report

R11922131 資工碩一 周柏諺

==========

Part1.

1.今天有一個能夠分類真假新聞的模型，而當這個模型被攻擊的話可能會導致假訊息被分類為真的訊息，此時若有某些人需要根據這方面的消息做出決策時，就會因此導致他們誤判做出不利的選擇。舉例來說，若有假消息是關於一間公司即將倒閉，但因為被誤分類為真的消息，而在人們誤信這則消息後，造成其股價下跌，許多人的錢都損失了。

2.因為 NLP 的 input 是離散的，在 cv 的 attack 可以透過對 image 加入 noise，但在 nlp 中，只能對離散的 token 攻擊，無法對 word embedding 進行攻擊。

3.

Goal : what the attack aims to achieve

Transformations : how to construct perturbations for possible adversaries,

Constrains : what a vaild adversarial example should satisfy

search method : how to find an adversarial example from the transformations that satisfies the onstrains and meets the goal

4.

Goal : untargeted classification

Constrains : word embedding distance, USE sentence similarity, POS consistency

Transformation : word substitution by counter-fitted Glove embedding space

Search method : greedy search with word importance ranking

主要先計算出每個字的 importance score，再過濾掉 stop words(ex:it, they, the …)，接著透過同義詞變換等攻擊的方式，並且對於替換的字要滿足一些限制，因為有做 counter-fitted，所以相同意思的詞會落在較接近的空間中，使得替換前後句子的相似度不能差太多，利用這種攻擊手法使得 model 誤判。

Part2.

1. yes ,cat.

2. no

3. d. JPEG compression reduces the noise level.