ADL—GAN

2019年4月30日　　上午 09:15

D輸出scalar，越真就越大
先fix住G，先train D
D會從真的跟假的去判別是1還是0(分類regression)
在固定住D，讓G騙過D
給G一個向量生成圖片，讓D的輸出值要越大越好

G跟D接再一起變成比較大比較深的NN
最後幾個layer假裝是D
中間會有一個特別寬的hidden layer當作是圖片
可以去訓練G用gradient ascend
=
生成人臉:
主要作用是可以產生臉的連續變化，不會產生雙面臉，而是可以自己學到轉向，不是所有向量疊加

(Variational)Auto-Encoder:
拿decoder出來來做生成
AE是為了產生現有資料庫的某一張圖片，所以就會是訓練資料的某一張，希望產生出來的片跟某一張資料庫圖片越接近越好，那因為兩張圖片色塊相近的話就會loss很小了，但人眼看就會發現很不一樣

GAN的D沒看過資料庫的圖片，所以不是背好的，而是用更一般化的方法看圖片裡面，G也會因此學到更好的general特徵
D是一個sigmoid binary cross entropy

GAN的結果大好大壞，VAE很穩定結果，但是最好的結果還是GAN

從prior distribution sample很多vector，產生圖片x~，去跟真實圖片算sigmoid binary cross entropy。把D train一些steps，只能找到local lower bound，量出JS divergence(update很多次)
G要去minize這個式子，G不能train太多，否則會無法evaluate，只會update一次就好，讓JS divergence變小。

- 但因為太難train (一開始loss太小)，所以從log(1-D(x))改成-log((D(x)))
- 把binary clf的label換過來，把G的label變成1，把真實的label改成0=> NSGAN (non S GAN)
- 原始的理論loss叫做MMGAN (min max GAN)

=理論=
輸入簡單的distribution z，經過G，產生複雜x的distribution，希望產生的分

布越接近資料庫所有圖片的distribution
用Divergence衡量分布的差異，希望G對真實data的divergence越接近越好
怎麼衡量distribution的差距?

從這兩個distribution sample各自data，去抽出圖片出來，sample data point出來，給很多random vector給G
sample出各自data(G跟real)，去衡量Pg跟Pdata，用D去衡量
用data衡量
調整D讓某個式子越大越好

- D看到從real出來的越大越好，看到Pg給越低的分數(D(x))
- binary crossentropy也是這樣
- JS divergence，這兩個distribution有多不同
- 如果divergence差異小，maximun value會小。
- min-max problem (nash equivalent)
- 實際上寫loss 不是這樣照理論寫，可以參考FGAN

loss大代表JS divergence小，當兩者的distribution一樣的時候，D會壞掉，每一個地方出現的機率會都一樣

D train到後來不一定會壞掉，可能是反映兩個data distribution的divergence
sample G的圖片的時候，如果多拿了過去G產生的東西也當成fake data會表現比較好
所以D在做的也未必是在衡量distribution而已

很難訓練

- 因為Pg跟Pdata是沒有重疊的
  - 圖片是高維空間中低維的manifold。(平面中的兩條線重疊很小)
  - Pg跟Pdata是從真實distribution做sample，只有sample的結果，根據sample結果可能是沒有重疊的，因為圖片可能是不一樣的
  - 沒有重疊的JS divergence是會有問題的，不管距離多遠算出來的divergence永遠都是log2，但實際上有好壞之分
    - 對D而言永遠都很好分，acc都是100%
- Wasserstein distance
  - 把土推到某一個地方的距離就是d (從P移動到Q)
  - 有不只一種鏟土的方法，所以distance可能會不同
    - 所以算法是窮舉各種不同的鏟土方法，看哪一種方法最短，當成距離
    - 要解optimize problem
  - 所以變成不同距離的兩條線會有不一樣的距離
  - 改了loss function就變成WGAN，Pdata越高越好，Pg分數越低越好
  - D要是1-Lipschitz function，D要夠平滑，不然不會收斂
    - 不能給Pdata分數都是無限大，Pg都是-無限大，中間gap會太大，所以要加限制

    - weight clipping
    - Improved WGAN (gradient penalty)
    - Improved Improved WGAN
    - Spectral Normalization，讓某些地方的gradient = 1

Tip:
train的時候從normal distribution sample，test的時候從variance更小variance的distribution
=> mode collapse問題，產生都一樣的圖片
=> mode dropping，沒辦法產生各種膚色的人臉
解法=>用ensemble G，各自產生一張圖片

==Conditional Generation==
- 控制產生的東西
- 給G一個condition，產生要生成的東西，希望用文字產生他
- 一般文字轉image需要pair data，supervised learning
    - 產生圖片希望跟輸入文字對應的圖片越近越好，L1或L2 distance會變成火車的平均
- conditional GAN
    - 輸入文字跟影像
    - 吃一張圖片越像越好，給D一段文字敘述跟圖片給出分數
        - 圖片多真實，放在一起多匹配
        - output分數要帶有兩個意思，對應關係好給高分
        - 文字敘述跟產生的圖片pair不好要給低分，圖片不好要給低分 (兩種狀況)
    - input兩個東西(domain type會不一樣)，所以先用兩種network壓成一樣的embedding，把兩種意思拆解開來
        - 給出圖片x看夠不夠真
        - 再把圖片x跟condition c看有沒有pair起來
    - 收集pair data
- Image translation pix2pix
    - 有image pair data，幾何圖案變成真實房子
    - 如果只用supervised會很模糊
    - 所以D要看condition跟G的output來判斷好或不好。但可能會無中生有產生本來不想要的東西
        - 所以可以再接回supervised loss
- Video Generation
    - 給G video前半段，預測下個畫面
- Sound to image
    - 聲音跟影像的對應關係
    - 一小段聲音訊號，可以產生圖片
- Image to label
    - multi label image classifiction，要看出圖片有什麼東西

- 把clf當成condition G，想要產生圖片有什麼東西
- 用conditional GAN比superised還要好

- Unsupervised Conditional GAN
  - 沒有對應關係
  - 可以學一個G，把X domain轉乘Y
  1. 訓練G給他X domain的東西轉成Y domain
  2. 要產生中間產物
  - 第一種作法是cycle GAN
    - network不要太深就不會轉換不好
    - 把G的輸入跟輸出都訂進去pre-train好的encoder network
      EX:VGG希望得到結果越近越好
    - cycle GAN
      - 希望輸入跟輸出越接近越好
      - 第一個G要產生的東西要夠雲本，才可以透過第二個G產生的圖片還原 (cycle consistency)
      - 雙向的
      - disco GAN，dual GAN
      - biGAN = ALI
  - 第二種方法是訓練一群encoder跟decoder
    - 輸入一張X domain的圖，把最重要的東西抽出來。經過encoder期望畫出Y domain的東西
    - 分開訓練兩個domain的autoencoder
    - 在訓練分別domain的Discriminator
    - 為了讓Y domain的decoder要可以看懂X domain encoder的code，講同樣語言的話
      - Couple GAN、UNIT
        - 讓encoder、decoder有一些參數是一樣的
      - 在train一個discriminator
        - 希望可以不同domain聯手騙過domain discriminator，期望他們不同維度代表的是一樣的
      - apply cycle consistency
        - Combo GAN(類似cycle GAN)
      - Semantic Consistency
        - XGAN，希望latent represntation要一樣

=GAN用在文字跟語言上應用=

- 文字的style transfer
  - 兩堆不同風格間的轉換
    - 正面變成負面，負面變成正面
  - 可以用cycle GAN
  - Discete issue
    - 因為輸入跟輸出的文字長度不一樣
    - G可能會是seq2seq會變成離散東西的network，所以無法BP

        - 把word sequence轉成vector (還有很多其他方法)
- 文章轉成簡短摘要
    - 用seq2seq需要百萬篇以上的paired data
    - 希望可以給他一大堆文章給他一大堆摘要，不需要paired 關係
    - 給G讀一篇文章產生文字，訓練D看了很多摘要，希望G產生的東西
    - 第二個G，要看了摘要還原回原本文章，所以需要包含文章的重要資訊
    - seq2seq2seq auto-encoder，希望可以解回來，但中間的latent representation也要人可以看得懂，所以需要D看懂那些暗號
- Voice Conversion
    - 昔日要兩個人講一樣的聲音
    - 希望兩個人不需要講一樣的句子
    - cycle GAN
    - AE domain互換，讓不同說話的人的encoder都一樣，希望可以抽出重要資訊，universal encoder
        - universal decoder，但要和不同聲音，就又有不一樣的input
        - 最簡單的方法，就是把它講的那一句話抽出他的與者vector
    - 需要D可以確保不同domain的聲音是一樣的，讓vector只有內容的資訊沒有與者資訊
    - 一大堆與者，有一個encoder把資訊吃進去，要去騙過D，encdoer為了要騙過D不知道是誰講得所以只保留phoneing資訊
    - 然後把與者的聲音加進去，希望decoder產生句子要跟原本的與者說的話一樣
    - testing的時候就用emcdoer抽出資訊，加上某個與者的code產生新的聲音說同一句話
        - 會有train test mismatch問題
        - 所以在訓練的時候加上另一個D，希望可以判斷產生的聲音是真正的聲音，還有另一個D產生的聲音訊號要知道是輸入的code的那個語者講的
- Unsupervised Conditional Generation
    - unsupervised speech recognition
        - 一般supervised要收集很多語音文字pair，可以辨識英文
        - 有很多很少人用的語言
        - 希望機器可以用很少的label只聽了文字跟語音就可以得到辨識
        - 只有機器再聽大家在講話
        - 自己學會語音辨識 ，聽了大量語言語音辨識
    - 從語音找到pattern，不知道聲音訊號編號是什麼，希望可以把編號轉成文字
        - 把聲音訊號變成聲音token，人看不懂token是什麼
        - 把英文也轉成phoneme

- 利用cycle GAN

=REVIEW CONCLUSION=
D輸出scalar，越真就越大
先fix住G，先train D
D會從真的跟假的去判別是1還是0(分類regression)
在固定住D，讓G騙過D
給G一個向量生成圖片，讓D的輸出值要越大越好

G跟D接再一起變成比較大比較深的NN
最後幾個layer假裝是D
中間會有一個特別寬的hidden layer當作是圖片
可以去訓練G用gradient ascend
=
生成人臉:
主要作用是可以產生臉的連續變化，不會產生雙面臉，而是可以自己學到轉向，不是所有向量疊加

(Variational)Auto-Encoder:
拿decoder出來來做生成
AE是為了產生現有資料庫的某一張圖片，所以就會是訓練資料的某一張，希望產生出來的片跟某一張資料庫圖片越接近越好，那因為兩張圖片色塊相近的話就會loss很小了，但人眼看就會發現很不一樣

GAN的D沒看過資料庫的圖片，所以不是背好的，而是用更一般化的方法看圖片裡面，G也會因此學到更好的general特徵
D是一個sigmoid binary cross entropy

GAN的結果大好大壞，VAE很穩定結果，但是最好的結果還是GAN

從prior distribution sample很多vector，產生圖片x~，去跟真實圖片算sigmoid binary cross entropy。把D train一些steps，只能找到local lower bound，量出JS divergence(update很多次)
G要去minize這個式子，G不能train太多，否則會無法evaluate，只會update一次就好，讓JS divergence變小。

- 但因為太難train (一開始loss太小)，所以從log(1-D(x))改成-log((D(x)))
- 把binary clf的label換過來，把G的label變成1，把真實的label改成0=> NSGAN (non S GAN)
- 原始的理論loss叫做MMGAN (min max GAN)

=理論=
輸入簡單的distribution z，經過G，產生複雜x的distribution，希望產生的分布越接近資料庫所有圖片的distribution
用Divergence衡量分布的差異，希望G對真實data的divergence越接近越好
怎麼衡量distribution的差距?

從這兩個distribution sample各自data，去抽出圖片出來，sample data point出來，給很多random vector給G
sample出各自data(G跟real)，去衡量Pg跟Pdata，用D去衡量
用data衡量
調整D讓某個式子越大越好

- D看到從real出來的越大越好，看到Pg給越低的分數(D(x))
- binary crossentropy也是這樣
- JS divergence，這兩個distribution有多不同
- 如果divergence差異小，maximun value會小。
- min-max problem (nash equivalent)
- 實際上寫loss 不是這樣照理論寫，可以參考FGAN

loss大代表JS divergence小，當兩者的distribution一樣的時候，D會壞掉，每一個地方出現的機率會都一樣

D train到後來不一定會壞掉，可能是反映兩個data distribution的divergence sample G的圖片的時候，如果多拿了過去G產生的東西也當成fake data會表現比較好
所以D在做的也未必是在衡量distribution而已

很難訓練

- 因為Pg跟Pdata是沒有重疊的
  - 圖片是高維空間中低維的manifold。(平面中的兩條線重疊很小)
  - Pg跟Pdata是從真實distribution做sample，只有sample的結果，根據sample結果可能是沒有重疊的，因為圖片可能是不一樣的
  - 沒有重疊的JS divergence是會有問題的，不管距離多遠算出來的divergence永遠都是log2，但實際上有好壞之分
    - 對D而言永遠都很好分，acc都是100%
- Wasserstein distance
  - 把土推到某一個地方的距離就是d (從P移動到Q)
  - 有不只一種鏟土的方法，所以distance可能會不同
    - 所以算法是窮舉各種不同的鏟土方法，看哪一種方法最短，當成距離
    - 要解optimize problem
  - 所以變成不同距離的兩條線會有不一樣的距離
  - 改了loss function就變成WGAN，Pdata越高越好，Pg分數越低越好
  - D要是1-Lipschitz function，D要夠平滑，不然不會收斂
    - 不能給Pdata分數都是無限大，Pg都是-無限大，中間gap會太大，所以要加限制
  - weight clipping
  - Improved WGAN (gradient penalty)
  - Improved Improved WGAN

- Spectral Normalization，讓某些地方的gradient = 1

Tip:
train的時候從normal distribution sample，test的時候從variance更小variance的distribution
=> mode collapse問題，產生都一樣的圖片
=> mode dropping，沒辦法產生各種膚色的人臉
解法=>用ensemble G，各自產生一張圖片

==Conditional Generation==
- 控制產生的東西
- 給G一個condition，產生要生成的東西，希望用文字產生他
- 一般文字轉image需要pair data，supervised learning
  - 產生圖片希望跟輸入文字對應的圖片越近越好，L1或L2 distance會變成火車的平均
- conditional GAN
  - 輸入文字跟影像
  - 吃一張圖片越像越好，給D一段文字敘述跟圖片給出分數
    - 圖片多真實，放在一起多匹配
    - output分數要帶有兩個意思，對應關係好給高分
    - 文字敘述跟產生的圖片pair不好要給低分，圖片不好要給低分 (兩種狀況)
  - input兩個東西(domain type會不一樣)，所以先用兩種network壓成一樣的embedding，把兩種意思拆解開來
    - 給出圖片x看夠不夠真
    - 再把圖片x跟condition c看有沒有pair起來
  - 收集pair data
- Image translation pix2pix
  - 有image pair data，幾何圖案變成真實房子
  - 如果只用supervised會很模糊
  - 所以D要看condition跟G的output來判斷好或不好。但可能會無中生有產生本來不想要的東西
    - 所以可以再接回supervised loss
- Video Generation
  - 給G video前半段，預測下個畫面
- Sound to image
  - 聲音跟影像的對應關係
  - 一小段聲音訊號，可以產生圖片
- Image to label
  - multi label image classifiction，要看出圖片有什麼東西
  - 把clf當成condition G，想要產生圖片有什麼東西
  - 用conditional GAN比superised還要好
- Unsupervised Conditional GAN

- 沒有對應關係
- 可以學一個G，把X domain轉乘Y

1. 訓練G給他X domain的東西轉成Y domain
2. 要產生中間產物

- 第一種作法是cycle GAN
  - network不要太深就不會轉換不好
  - 把G的輸入跟輸出都訂進去pre-train好的encoder network
    EX:VGG希望得到結果越近越好
  - cycle GAN
    - 希望輸入跟輸出越接近越好
    - 第一個G要產生的東西要夠雲本，才可以透過第二個G產生的圖片還原 (cycle consistency)
    - 雙向的
    - disco GAN，dual GAN
    - biGAN = ALI
- 第二種方法是訓練一群encoder跟decoder
  - 輸入一張X domain的圖，把最重要的東西抽出來。經過encoder期望畫出Y domain的東西
  - 分開訓練兩個domain的autoencoder
  - 在訓練分別domain的Discriminator
  - 為了讓Y domain的decoder要可以看懂X domain encoder的code，講同樣語言的話
    - Couple GAN、UNIT
      - 讓encoder、decoder有一些參數是一樣的
    - 在train一個discriminator
      - 希望可以不同domain聯手騙過domain discriminator，期望他們不同維度代表的是一樣的
    - apply cycle consistency
      - Combo GAN(類似cycle GAN)
    - Semantic Consistency
      - XGAN，希望latent represntation要一樣

=GAN用在文字跟語言上應用=

- 文字的style transfer
  - 兩堆不同風格間的轉換
    - 正面變成負面，負面變成正面
  - 可以用cycle GAN