1 Introduction

Il paper in questione analizza un nuovo sistema chiamato *CAN (Creative Adversarial Netowrks)*. Quest'ultimo genera arte guardando in generale tra i vari movimenti artistici e in particolare ne apprende lo stile. Tale modello è una modifica del GAN (Generative Adversaria Networks) la cui abilità è quella di apprendere e generare immagini nuove simulando una data distribuzione. In generale i modelli sono limitati nel creare prodotti creativi nel loro design originale. L'obiettivo principale dunque è quello di realizzare un sistema creativo per la generazione dell'arte senza l'aiuto di artisti nel processo creativo.
Il processo di creazione da parte dell'uomo utilizza una *prior experience (conoscenza passata)* e l'esposizione all'arte. Colin Martindale ha ipotizzato che gli artisti creativi provano ad incrementare il *potenziale di eccitazione - arousal potential -* (la capacità d'eccitazione da parte di uno stimolo) in modo che quest'ultimo venga spinto verso il concetto di *abitudine (in senso artistico)*. E' stato dimostrato da *Wundt* che il gradimento è inversamente proporzionale allo stimolo. In particolare un essere umano ha un basso potenziale di eccitazione quando è rilassato, mentre è alto quando quest'ultimo è presente in una situazione passionale, violenta o per esempio quando è infuriato. Solitamente le persone identificano la nuova arte generata da un sistema automatico come psichedelica; tale reazione negativa può essere spiegata come risultato del troppo incremento del potenziale di eccitazione.
A partire da questa teoria analizziamo quindi le modifiche necessarie alla realizzazione del nuovo sistema che permetterà la generazione di arte creativa massimizzando la deviazione da stili prestabiliti e minimizzandola a partire da una distribuzione sull'arte. Nel sistema GAN il generatore prova soltanto a generare immagini simili a quelle di training. Dopo un'accurato training, il discriminatore non dovrebbe più essere in grado di capire se l'immagine è reale o fake (dunque il generatore ha appreso la distribuzione originale del training set ricevendo di volta in volta un segnale da parte del discriminatore). Il problema di questo modello quindi è che non viene creata nuova arte, ma soltanto emulata a partire da quella già presente.
Nel modello attuale, il generatore riceve due feedback contraddittori in modo da realizzare tre obiettivi: generare nuovi lavori (ma non troppo nuovi), non deve allontanarsi troppo dalla distribuzione originale o aumentare troppo il potenziale d'eccitazione e infine generare nuove "opere" che permettono l'incremento dell'ambiguità stilistica. Come in precedenza il CAN ha due reti avversarie: il discriminatore e il generatore. Rispetto a prima quì il generatore riceve due segnali dal discriminatore: immagine reale o no e quanto bene il discriminatore riesce a classificare il nuovo lavoro generato in un particolare stile. Il generatore quindi ha il compito di generare nuova arte in grado di confondere il discriminatore (indurre a pensare che sia effettivamente arte) ma non solo, il generatore ha il compito di confonderlo anche relativo allo stile.
In precedenza abbiamo detto che i signali inviati da parte del discriminatore sono contraddittori: il primo segnale (arte/non arte) forza il generatore a generare nuove immagini in modo tale che il discriminatore accetta come arte. Se riesce entro le regole degli stili stabiliti, il discriminatore sarà anche in grado di classificarne lo stile.

1.1 Training and Architecture

Facendo un paragone con il GAN, il quale è tipicamente allenato come un gioco tra due giocatori, anche il CAN viene allenato allo stesso modo. Tipicamente i due avversari sono modellati come reti neurali, utilizzando un *min-max game* per il training. Per la realizzazione del CAN viene modificata la loss del GAN aggiungendo la *style classification loss* e la *style ambiguity loss*. La massimizzazione dell'ambiguità di stile può essere ottenuta massimizzando la *style class posterior entropy*, costruendo una loss function in modo tale che il generatore G produce un'immagine $x \sim p_{data}$ nel frattempo che massimizza $p(c|x)$ (style class posterior) per l'immagine generata. Se utilizzassimo l'entropia, quest'ultima è massimizzata quando le classi a posteriori $p(c|G(z))$ sono equiprobabili. Se invece utilizzassimo la cross-entropy (differenza di similarità tra due distribuzioni) con distribuzione uniforme, quest'ultima sarà minimizzata quando le classi sono equiprobabili. Entrambi gli obiettivi quindi saranno ottimali quando le classi sono equiprobabili. L'utlizzo della cross-entropy risulta pesantemente penalizzata per samples classificati correttamente. Ciò comporta che in alcuni casi si avrà una loss grande, che comporta di conseguenza a enormi valori per il calcolo del gradiente. Definiamo quindi la funzione di costo con differenti obiettivi avversari:

$$min_G max_D V(D,G) = \mathbb{E}_{x,\hat{c}\sim p_{data}}[logD_r(x) + logD_c(c=\hat{c}|x)]+$$

$$\mathbb{E}_{z\sim p_z}[log(1 - Dr(G(z))) - \sum_{k=1}^{K}(\frac{1}{K}log(D_c(c_k|G(z))+$$

$$(1 - \frac{1}{K}log(1 - D_c(c_k|G(z)))]$$

dove z (noisy vector) $\sim p_z \sim N(\mu, \sigma^2)$, $(x = real - image, \hat{c} = style) \sim p_{data}$, $D_r(.) =$ *(transformation function) funzione che discrimina tra immagini reali e generate* e $D_c(.) =$ funzione che discrimina tra differenti stili e *stima la posterior style class* $(D_c(c_k|.) = p(c_k|.))$. Ottimizzando G e D in modo alternato il modello sarà in grado di generare immagini che emulano la distribuzione di training.

1.1.1 Discriminator Training

Minimizzando $-\mathbb{E}_{x\sim p_{data}}[logD_r(x)+logD_c(c=\hat{c}|x)]$ per le immagini reali e $-\mathbb{E}_{z\sim p_z}[log(1-Dr(G(z)))]$ si massimizza l'equazione in precedenza. Il discriminatore quindi non si allena soltanto a identificare l'arte reale, ma anche ad indentificare la loro classe di stile attraverso la *K-way loss*. Inoltre, il discriminatore simulataneamente apprende sia la distribuzione dell'arte che dei vari stili.

1.1.2 Generator Training

Caso opposto, il generatore tende a minimizzare l'equazione precedente. Questo avviene massimizzando $log(1-Dr(G(z)))-\sum_{k=1}^{K}(\frac{1}{K}log(D_c(c_k|G(z))+(1-\frac{1}{K}log(1-D_c(c_k|G(z)))$. In questo modo le immagini generate assomigliano a quelle reali (primo termine) e nel

frattempo si ha una grande *cross-entropy* per $p(c|G(z))$ (class posterior) avente distribuzione uniforme per massimizzare la *style ambiguity* (secondo termine). Come possiamo notare, in questo modello non si utilizza nessuna *class label* come nei generative model.

1.1.3 Model Architecture

Il generatore G prende $z \in \mathbb{R}^{100}$ (assume valori da 0 a 1) e lo estende 4 volte di più nello spazio convoluzionale rappresentato con 2048 feature maps. Una serie di quattro fractionally strided convolutions, riferite erroneamente come deconvolutions, traspongono l'immagine, per poterla trasformare da un formato piccolo ad uno grande. Per far questo, un fractionally strided convolution ricostruisce lo spazio di risoluzione dell'immagine eseguendo dopo l'operazione di convoluzione. La rappresentazione ad alto livello infine viene convertita in un'immagine 256x256. Quindi si parte da $z \in \mathbb{R}^{100} \longrightarrow 4x4x1024 \longrightarrow 8x8x1024 \longrightarrow 16x16x512 \longrightarrow 32x32x256 \longrightarrow 64x64x128 \longrightarrow 128x128x64 \longrightarrow 256x256x3$ (la dimensione dell'immagine originale).
Il discriminatore ha un corpo comune di layer convoluzionali seguiti da due teste (una per il real/fake, una per multi-label loss). Il *corpo comune* di layer convoluzionali è composta da una serie di layer convoluzionali (stride=2, padding=1). Inoltre, ciascun layer ha come funzione di attivazione non lineare la *LeakyRelU*

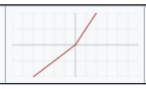
Leaky rectified linear unit (Leaky ReLU)[17]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$

Figure 1: LeakyRelU

Dopo aver passato l'immagine al corpo comune D, questo produrrà una feature map di (4x4x512). Il discriminatore D_r collassa il tensore $4x4x512$ in un layer denso producendo $D_r(c|x)$ ovvero la probabilità che un'immagine venga dalla distribuzione delle immagini reali. La *multi-label probabilities* $D_c(c_k|x)$ è prodotta passando la feature maps di dimensione $4x4x512$ in 3 layer densi di dimensioni 1024,512 e K.

1.1.3.0.1 Initialization and Training I pesi sono inizializzati da una distribuzione normale centrata in 0 con una deviazione standard pari a 0.02. Si usa il batch size pari a 128 con il mini-batch SGD per il training con 0.0001 come learning rate e 100 epoche (100 volte visualizzato l'intero set di training). Per stabilizzare il training si usa la Batch Normalization che lo normalizza per avere $\sim N(0, 1)$. Inoltre si esegue la *data agumentation* dove viene preso il 90

1.2 Qualitative Validation

Valutare la creatività di un artefatto generato da una macchina è un problema davvero complesso. Per la valutazione si utilizzano immagini generate da 3 modelli base:
•Il primo modello è l'originale DCGAN di Google. Questo modello è capace di generare immagini grandi 64x64, sebbene però quest'ultimo è allenato sul dataset, quest'ultimo

fallisce nell'emulare l'arte, poichè le immagini generate non mostrano alcuna figura o generi/stili artistici.
•Il secondo è una variazione del precedente in cui l'immagine generate ha dimensione 256x256. Il generatore in più ha la stessa architettura del CAN. Nei risultati possiamo vedere come si hanno miglioramenti significativi; in particolare le immagini sono esteticamente accattivanti nella struttura e nei contrasti tra i vari colori.
•Il terzo modello (style-classification-CAN) è una modifica del CAN, in cui viene soltanto considerata la style classification loss (viene tolta la style ambiguity loss). In questo modello il discriminatore apprende la discriminazione tra classi di stile attraverso l'apprendimento della distribuzione dell'arte. Il generatore quindi ha la stessa loss come nel GAN. Si ha un significativo incremento nell'emulazione della distribuzione dell'arte in termini di sfondi, architetture, figure religiose ecc.. Ciò non accade nei due modelli precedenti.
I primi due modelli dunque apprendono ciò che è arte o ciò che non lo è. Mentre l'ultimo può differenziare anche lo stile di tale arte.
•**IMPORTANTE:** Ciò che stiamo dimostrando quindi è che il style-classification-CAN model può emulare maggiormente la distribuzione dell'arte apprendendo anche il loro stile, con l'unico difetto che non è creativo (il nostro obiettivo.
Il modello CAN invece genera immagini che possono essere caratterizzate come nuove e che non emulano alcuna distribuzione dell'arte rendendole comunque esteticamente accattivanti. Il problema è che quest'ultimo non mostra figure tipiche, generi, stili o particolari soggetti (simile ai primi due modelli citati in precedenza). Non possiamo affermare che ciò accada perchè non riesce a simulare la distribuzione dell'arte poichè semplicemente togliendo la *style loss ambiguity* (modello style-classification-CAN) il quale genera in modo accattivante tali elementi. Ciò che possiamo affermare quindi è che la *style ambiguity loss* forza il modello a generare nuove immagini e, allo stesso tempo, rimane vicina alla distribuzione dell'arte (il che rende l'immagine generata accattivante). Cerchiamo di capirlo meglio con un esperimento quantitativo.

1.3 Quantitative Validation

L'obiettivo degli esperimenti che vedremo è quello di capire se un essere umano è capace di distinguere l'arte generata da un umano vs quella generata da una macchina. Poichè ne vogliamo valutare la creatività degli artefatti prodotti dal sistema, dobbiamo comparare le risposte ricevute alla sola arte che è considerata nuova e creativa in questo punto del tempo (se valutassimo opere impressioniste stiamo valutando la capacità di emulare tali opere, e non di essere creativi). Per questo motivo sono stati creati due set per i capolavori degli artisti reali, e quattro per quelli generati dalla macchina:
•1 (25 opere). Espressionismo astratto: utilizzato poichè mancano figure o soggetti in chiaro. L'esistenza di figure o soggetti potrebbe influenzare la decisione dell'osservatore che le opere sono fatte da un uomo poichè una macchina le immagini generate che mancano di tali figure.
•2 (25 opere). Arte contemporanea: Questo set è composto da 25 dipinti di vari artisti che sono stati esposti in Art Basel 2016. Si tratta di opere d'arte ai confini della creatività umana nella pittura, almeno come giudicato dagli esperti d'arte e dal mercato

dell'arte. Il set è selezionato a caso dopo aver escluso l'arte che ha figure chiare o evidenti pennellate che potrebbero disturbare i soggetti.
•3/4 (100-76 immagini): Immagini generate da DCGAN risoluzione 64x64 e 256x256.
•5 (100 immagini): immagini generate da sc-CAN. Usiamo queste immagini per comparare l'efficacia nell'aggiunta della *style ambiguity loss*.
•6 (125): Generate dal CAN.

1.3.1 Experiment 1

L'obiettivo di questo esperimento è testare la capacità del sistema di generare arte che gli essere umani non sono in grado di distinguere dalla migliore arte creativa che viene generata al giorno d'oggi dagli artisti. Ad ogni soggetto viene presentata un'immagine alla volta e quest'ultimo dovrà rispondere a due domande: dovrà scegliere se l'immagine che sta vedendo è generata dall'uomo o da un macchina, e deve valutarla con un voto da 1 a cinque. L'esperimento mostra che i vari utenti classificano le imamgini del set "Epressionismo Astratto" come generate da un'artista (85%). Il modello CAN proposto ha sovraperformato il modello GAN generando immagini che i soggetti umani pensano siano generati dall'artista. Sono stati eseguiti due test su due campioni per determinare il significato statistico di questi risultati, con l'ipotesi nulla che le risposte dei soggetti sia per CAN che per GAN provengano dalla stessa distribuzione. Il t-test (è un test statistico di tipo parametrico con lo scopo di verificare se il valore medio di una distribuzione si discosta significativamente da un certo valore di riferimento) ha respinto tale ipotesi con diversi valori di p-value (il valore p aiuta a capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità introdotta dal campionamento, oppure se tale differenza è statisticamente significativa, cioè difficilmente spiegabile mediante la casualità dovuta al campionamento.)(TABELLA 2). Ovviamente non possiamo dire che il CAN è più creativo del GAN. Infatti un sistema che copia perfettamente l'arte umana, senza essere innovativo, otterrebbe un punteggio più alto in queste domande. Tuttavia, possiamo escludere questa possibilità poiché le immagini generate da CAN e GAN non copiano l'arte umana come è stato spiegato precedentemente.

1.3.2 Experiment 2

In questo esperimento vengono proposte prima una serie di domande e poi alla fine viene chiesto se l'immagine mostrata è generata da un artista o da una macchina. è stato ipotizzato che se questa domanda venisse posta per prima agli utenti dell'esperimento, quest'ultimi avrebbero maggiori probabilità di rispondere in modo casuale. Mentre attraverso una serie di domande la risposta potrebbe essere più costruttiva.
Gli esperimenti mostrano come dopo una serie di domande, le risposte all'ultima cambiano radicalemente da set a set. I risultati mostrano come CAN genera immagini migliori rispetto a GAN (75% vs 65%) (TABELLA 3). Sebbene abbiamo utilizzato lo stesso dataset si ha un'incremento significativo nelle risposte del modello CAN (75 % vs 53 %).

1.3.3 Exmperiment 3

In questo esperimento viene giudicato se l'arte generata dal CAN può essere considerata davvero arte. L'esperimento è molto simile a quello di *Snapper* per determinare se un lavoro è intenzionale, ha una struttura visuale, è comunicativo e ispirativo. L'ipotesi è quella che un umano preferirà le opere di artisti reali, basandosi su queste particolari domande, rispetto a quelle generate dalla macchina. I risultati però mostrano ben altro, difatti vengono preferite quelle della macchina anzicchè quelle degli artisti. Il fatto che i soggetti abbiano trovato le immagini generate dalla macchina intenzionalmente, strutturate visivamente, comunicative e stimolanti, con livelli simili all'arte umana reale, indica che i soggetti vedono queste immagini come arte.

1.3.4 Experiment 4

L'obiettivo di questo esperimento è valutare l'effetto dell'aggiunta della perdita di ambiguità di stile al modello CAN, in contrasto con la perdita di classificazione dello stile, nel generare immagini nuove ed esteticamente accattivanti. In altre parole, è conoscere gli stili o deviare dallo stile che rende i risultati creativi? Per valutare la creatività ci riferiamo alla definizione più comune di creatività di un artefatto: nuova e influente. Poichè l'influenza non è rilevante teniamo solo conto della novità per la metrica della creatività. Ad ogni soggetto sono state mostrate coppie di immagini, una dal modello CAN e una dal modello sc-CAN, selezionate casualmente e posizionate in ordine casuale fianco a fianco. I risultati di questo esperimento mostrano che il 59,47% delle volte i soggetti hanno selezionato CAN come più nuovo e il 60% delle volte hanno trovato le immagini di CAN più esteticamente attraenti. Ciò indica l'effetto del perdita di ambiguità di stile nel processo di generazione da parte del CAN comparato alla perdita di classificazione di stile.

1.4 Conclusion

Il sistema creato quindi è in grado di massimizzare l'ambiguità di stile nel frattempo che rimane vicino alla distribuzione dell'arte (classificazione di stile). L'interazione dei due segnali nel generatore permette la generazione di nuovi artefatti esplorando lo spazio creativo per trovare soluzioni che deviano dallo stile stabilito ma che rimangano quanto più possibile vicini al confine dell'arte per essere riconosciuta come tale. è bene però notare che il sistema non ha alcune conoscenza dell'arte in termini di concetti o stile. Non conosce alcun soggetto o elementi espliciti delle principali correnti artistiche. L'apprendimento è basato soltanto sull'esposizione dell'arte, concetti e stili. In questo modo il sistema ha l'abilità di apprendere sempre nuova arte e adattare (o almeno si spere) la sua generazione su ciò che apprende di nuovo.

1.5 Considerazioni Personali

Alla fine del paper sono state lasciate alcune domande aperte. Una in particolare ha attirato la mia attenzione ovvero: i soggetti sono influenzati da una estetica accattivante? E ciò che significa che i risultati non sono creativi?