

Trasferimento di Stile Neurale: Combinazione di Contenuto e Stile per la Generazione di Immagini Artistiche

Eugenio Quaglia

Abstract—L’obiettivo di questo paper è replicare e comprendere i risultati ottenuti da Gatys, Ecker e Bethge nel loro lavoro ”A Neural Algorithm of Artistic Style”. Utilizzando la rete neurale convoluzionale pre-addestrata VGG19, sviluppata dall’Università di Oxford per il riconoscimento delle immagini, abbiamo esplorato la capacità di separare e combinare il contenuto e lo stile di diverse immagini. In particolare, abbiamo generato nuove immagini che fondono il soggetto di una fotografia con lo stile di celebri dipinti, come ”La notte stellata” di Van Gogh, ”L’Urlo” di Munch e ”Convergence” di Pollock.

I. PROBLEM

”Come sarebbe Timothée Chalamet se vivesse all’interno di un dipinto di Van Gogh?” In questo paper, cercheremo di rispondere a questa domanda utilizzando VGG19, una rete neurale convoluzionale pre-addestrata, comunemente impiegata per la classificazione di immagini. Esploriamo ora l’architettura di VGG19.

La rete VGG-19 è una rete neurale convoluzionale sviluppata per il riconoscimento di oggetti e ampiamente descritta in studi precedenti. Questa rete possiede 19 layer totali, di cui 16 sono convoluzionali e 3 sono fully connected. Per il nostro studio, ci concentreremo solo sui layer convoluzionali, tralasciando i layer fully connected.

L’architettura della VGG-19 è basata su due calcoli fondamentali:

Convoluzione lineare rettificata: Utilizza filtri di dimensione $3 \times 3 \times k$, dove k è il numero di feature map di input. La convoluzione viene eseguita con uno stride e un padding pari a uno, in modo che la feature map di output abbia le stesse dimensioni spaziali delle feature map di input.

Max Pooling: Applica un pooling massimo su regioni non sovrapposte di dimensione 2×2 , riducendo le dimensioni spaziali delle feature map di un fattore di due.

Questi due calcoli sono applicati in modo alternato. Dopo un certo numero di layer convoluzionali, segue un layer di max pooling. Dopo ciascuno dei primi tre layer di pooling, il numero di feature map raddoppia. Insieme al downsampling spaziale, questa trasformazione porta a una riduzione del numero totale di risposte delle feature di un fattore di due.

Le immagini di input alla rete VGG-19 hanno una dimensione standard di 224×224 pixel. I layer convoluzionali utilizzano filtri con la dimensione minima possibile, ovvero 3×3 . Questo approccio ha un duplice vantaggio: uno stack di due layer convoluzionali 3×3 ha un campo di recezione effettivo di 5×5 (e uno stack di tre layer convoluzionali 3×3 ha un campo di recezione di 7×7 e così via), permettendo di applicare più volte una funzione di non linearità attraverso i layer convoluzionali. Inoltre, l’utilizzo di filtri 3×3 richiede

meno parametri da immagazzinare (ad esempio, $27C^2$ per tre filtri 3×3 rispetto a $49C^2$ per un filtro 7×7).

Dopo ogni layer convoluzionale, che applica una trasformazione lineare all’input, viene applicata una funzione non lineare, nel caso della VGG-19, si tratta della funzione ReLU. Tra un blocco convoluzionale e l’altro viene applicato un layer di max pooling che dimezza le dimensioni spaziali delle feature map. Grazie a questa struttura, i layer superiori della rete catturano il contenuto ad alto livello in termini di oggetti e la loro disposizione nell’immagine di input, mentre i layer inferiori riproducono dettagli più vicini ai valori dei pixel dell’immagine originale.

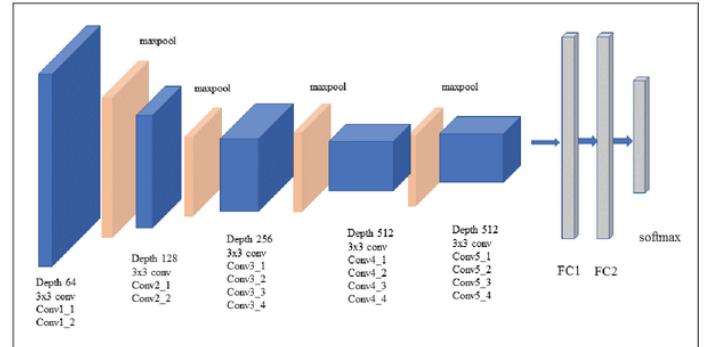


Fig. 3. VGG-19 network architecture

Fig. 1. Architettura della rete CNN Vgg19. Si può notare che tutte le convoluzioni sono 3×3 e che ci sono 5 stack di layer convoluzionali alternati a layer di max pooling. La parte fully connected non è utile ai nostri scopi

Quando addestriamo una rete convoluzionale su un oggetto, essa sviluppa una rappresentazione dell’immagine che rende le informazioni sugli oggetti sempre più esplicite lungo il processo di elaborazione. Man mano che si avanza nella gerarchia dei livelli, l’input viene trasformato in rappresentazioni che si concentrano progressivamente sul contenuto effettivo dell’immagine piuttosto che sui singoli pixel.

È possibile visualizzare direttamente le informazioni contenute in ciascun livello della rete ricostruendo l’immagine a partire solo dalle mappe di feature di quel livello. I livelli superiori della rete catturano il contenuto di un’immagine in termini di oggetti e della loro disposizione, senza vincolare i valori esatti della ricostruzione. Ci riferiremo a queste risposte delle feature nei livelli superiori come alla rappresentazione del contenuto.

Per ottenere una rappresentazione dello stile di un’immagine, utilizziamo uno spazio delle feature progettato per catturare informazioni sulla texture. Questo spazio viene costruito basandosi sulle risposte dei filtri a ciascun livello

della rete e consiste nelle correlazioni tra le diverse risposte dei filtri su tutta l'estensione spaziale delle mappe delle feature. Includendo le correlazioni delle feature di più livelli, otteniamo una rappresentazione dell'immagine di input che cattura le informazioni sulla sua texture. Si può osservare che la dimensione e la complessità delle strutture locali dell'immagine di input aumentano con la profondità del layer considerato. Ci riferiremo alla somma pesata degli stili di tutti i livelli come alla rappresentazione dello stile.

Uno degli obiettivi chiave di questo lavoro è dimostrare che è possibile separare la componente di stile da quella di contenuto. Per fare ciò, genereremo nuove immagini partendo da una foto del celebre attore americano Timothée Chalamet, mescolandola con gli stili di dipinti famosi, in particolare "Notte stellata" di Van Gogh. Variaremo i parametri coinvolti per osservare come l'immagine risultante cambi in base a essi.

Ovviamente, il contenuto e lo stile di un'immagine non possono essere completamente separati. Quando si sintetizza un'immagine che combina il contenuto di una foto con lo stile di un'altra, non esiste una soluzione che soddisfi perfettamente entrambi i vincoli contemporaneamente. La funzione di perdita che andremo a minimizzare sarà composta da due termini distinti, uno per il contenuto e uno per lo stile, rendendo così possibile decidere su quale dei due aspetti si desidera porre maggiore enfasi.

II. METHODS

Come descritto precedentemente, la rete VGG è costituita da 16 layer convoluzionali, 5 di pooling e 3 fully connected, i quali però non verranno considerati nella nostra analisi. Ogni layer della rete definisce una banca di filtri non lineari la cui complessità aumenta con la profondità del layer stesso. Quando un'immagine X viene elaborata dalla CNN, essa viene codificata a ciascun livello tramite le risposte dei filtri applicati a quell'immagine. Un layer con N_l filtri distinti genera N_l mappe delle feature, ciascuna con dimensione M_l (dove M_l è il prodotto tra la larghezza e l'altezza della mappa stessa).

La risposta di un layer l può essere rappresentata da una matrice F_l di dimensioni $N_l \times M_l$, dove l'elemento $F_{l,ij}$ rappresenta l'attivazione del filtro i -esimo nella posizione j del layer l .

Per visualizzare le informazioni codificate nei vari livelli della gerarchia della rete (come mostrato in Fig. 1, ricostruzioni del contenuto), eseguiamo un'ottimizzazione tramite discesa del gradiente su un'immagine di rumore bianco per trovare un'altra immagine che produca risposte delle feature simili a quelle dell'immagine originale. Se indichiamo con p l'immagine generata e con x quella originale, e con P_l e F_l le rispettive rappresentazioni al layer l , possiamo definire la funzione di perdita per il contenuto come:

$$L_{\text{content}}(p, l) = \frac{1}{2} \sum_{i,j} (F_{l,ij}^p - P_{l,ij}^p)^2$$

Analogamente, per ottenere una rappresentazione dello stile di un'immagine, costruiamo uno spazio delle feature che

cattura le correlazioni tra le diverse risposte dei filtri, calcolate sull'intera estensione spaziale dell'immagine di input. Queste correlazioni sono rappresentate dalla matrice di Gram

$$G_l \in R^{N_l \times N_l},$$

dove $G_{l,ij}$ è il prodotto scalare tra le mappe delle feature vettorizzate i e j nel livello l :

$$G_{l,ij} = \sum_k F_{l,ik} F_{l,jk}$$

Per generare una texture che corrisponda allo stile di una determinata immagine (come illustrato in Fig. 1, ricostruzioni dello stile), eseguiamo un'ottimizzazione tramite discesa del gradiente partendo da un'immagine di rumore bianco, cercando di ottenere un'altra immagine che riproduca la rappresentazione dello stile dell'immagine originale. Questo viene ottenuto minimizzando la distanza quadratica media tra gli elementi della matrice di Gram dell'immagine originale e quelli della matrice di Gram dell'immagine generata. Indicando con α l'immagine originale e con x l'immagine generata, e con A_l e G_l le rispettive rappresentazioni di stile nel livello l , il contributo di quel livello alla perdita totale sarà:

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{l,ij}^l - A_{l,ij}^l)^2$$

e la perdita totale sarà:

$$L_{\text{style}}(a, x) = \sum_{l=0}^L w_l E_l$$

In questo modo, possiamo comprendere come la rete neurale scomponga e ricostruisca le componenti di contenuto e stile di un'immagine, generando una nuova immagine che combini queste due caratteristiche secondo i parametri stabiliti. Per ottenere le immagini che combinano il contenuto di una fotografia con lo stile di un dipinto, minimizziamo congiuntamente la distanza di un'immagine di rumore bianco dalla rappresentazione del contenuto della fotografia in un livello della rete e dalla rappresentazione dello stile del dipinto in diversi livelli della CNN. Siano p la fotografia e a l'opera d'arte. La funzione di perdita che minimizziamo è:

$$L_{\text{total}}(p, a, x) = \alpha L_{\text{content}}(p, x) + \beta L_{\text{style}}(a, x)$$

dove α e β sono i fattori di ponderazione per la ricostruzione del contenuto e dello stile, rispettivamente.

III. RESULT

La prima analisi che intendiamo condurre riguarda la variazione dell'informazione relativa al contenuto in funzione della profondità della rete. A partire da un'immagine di rumore bianco, è stato applicato il metodo del gradient descent utilizzando come target la mappa delle feature del livello di interesse. Questo processo è stato eseguito per i primi layer convolutivi di ciascuno stack, che, facendo riferimento alla nomenclatura adottata nell'immagine 1, corrispondono a

$conv1_1$, $conv2_1$, $conv3_1$, $conv4_1$ e $conv5_1$. Per semplicità, nel resto del paper ometteremo il secondo indice. Sono stati utilizzati i parametri $\alpha = 1$ e $\beta = 0$, con un learning rate di $lr = 0.004$ ed è stato eseguito il gradient descent per 6000 epoche. I risultati ottenuti sono riportati nell'immagine sottostante.

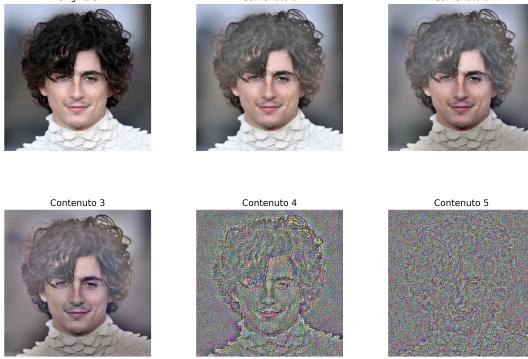


Fig. 2. Ricostruzioni delle immagini di contenuto per i vari layer convolutivi utilizzati nello studio. A contenuto 1 corrisponde conv 1, a contenuto 2 conv2 e così via

Nei primi strati della rete, la rappresentazione dell'immagine si focalizza su caratteristiche di basso livello, come bordi, angoli, texture semplici e dettagli locali. Questi strati sono responsabili del rilevamento di contorni, linee e piccoli dettagli presenti nell'immagine. In questa fase, l'immagine di contenuto risulta ancora relativamente simile a quella originale, sebbene venga leggermente modificata per enfatizzare queste caratteristiche fondamentali. Nei livelli più profondi, la rappresentazione dell'immagine si evolve verso concetti molto più astratti. In questi strati, la rete acquisisce la capacità di riconoscere oggetti complessi e caratteristiche globali dell'immagine. Ad esempio, la rete può identificare volti o oggetti specifici, piuttosto che semplici bordi o forme. A questo livello, l'immagine di contenuto appare significativamente diversa da quella originale, venendo rappresentata in modo più astratto, con informazioni utili per compiti come la classificazione o il riconoscimento.

È stata condotta un'analisi per esaminare come la rappresentazione dello stile vari al crescere della profondità dei layer della rete. Poiché la loss di stile è determinata dalla somma pesata dello stile di ciascun layer, l'analisi è stata eseguita partendo dal layer più superficiale e aggiungendo progressivamente i layer più profondi. Anche in questo caso, è emerso che, scendendo nella rete, la rappresentazione dello stile si evolve da una descrizione dettagliata di texture e pattern locali a una sintesi di caratteristiche stilistiche più globali e astratte. I primi strati catturano dettagli specifici, mentre i livelli più profondi si concentrano sull'essenza stilistica complessiva, integrando e astrattizzando le caratteristiche locali per produrre una visione più globale dello stile.

Le immagini sono state generate seguendo un procedimento

analogo a quello utilizzato per le immagini di contenuto; tuttavia, in questo caso, sono stati impiegati i parametri $\alpha = 0$, $\beta = 1$, un learning rate di $lr = 0.004$, 6000 epoche, e un vettore dei pesi dello stile $w_l = [1/5, 1/5, 1/5, 1/5, 1/5]$.

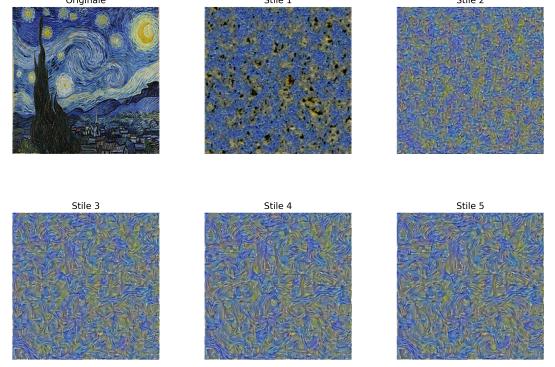


Fig. 3. Ricostruzioni delle immagini di stile per i vari layer convolutivi utilizzati nello studio. A stile 1 corrisponde lo stile di conv 1, a stile 2 lo stile di conv1 e conv2 e così via

Ora vogliamo esplorare i risultati ottenuti dalla combinazione di contenuto e stile. A tal fine, è stata utilizzata come immagine di contenuto una fotografia di Timothée Chalamet, mentre per l'immagine di stile sono stati scelti tre celebri dipinti: *La notte stellata* di Van Gogh, *L'urlo* di Munch e *Convergence* di Pollock.

Per realizzare questa fusione, sono stati impostati i seguenti parametri: $\alpha = 1$, $\beta = 0.05$, un learning rate di $lr = 0.004$, 6000 epoche, e un vettore dei pesi $w_l = [1, 0.75, 0.2, 0.2, 0.2]$. Il layer $conv_4$ è stato utilizzato come riferimento per lo stile, in linea con quanto proposto nel paper di Gatys. In questo caso, l'immagine di partenza per l'applicazione del gradient descent non è un'immagine di rumore bianco, ma una copia dell'immagine di contenuto. Questa scelta è stata fatta per ottimizzare il processo di minimizzazione della funzione di loss, migliorando l'efficienza e la convergenza dell'algoritmo.

I valori di questi parametri sono stati ispirati dal lavoro di Aneesh Aparajit su Kaggle, disponibile all'indirizzo link Kaggle.

È importante notare che il settaggio di questi parametri, in particolare il rapporto α/β e il vettore w_l , è altamente soggettivo e dipende dal risultato che si desidera ottenere. Per comprendere come l'immagine generata vari al cambiare dei valori di w_l , sono state prodotte due immagini: una enfatizzando lo stile del layer $conv_1$ e l'altra enfatizzando lo stile del layer $conv_5$.

Come verrà discusso nelle conclusioni, è complesso individuare un metodo oggettivo per valutare l'efficacia del nostro algoritmo. Le immagini generate possono essere considerate "buone" quando riescono a combinare in modo equilibrato le componenti di contenuto e stile; tuttavia, questo equilibrio è altamente soggettivo.



Fig. 4. Combinazione stile-contenuto tra una foto dell'attore Timothée Chalamet e un dipinto famoso, rispettivamente, da sinistra verso destra: "Notte stellata" di Van Gogh, "Sforzo d'uccello" di Pollock, "Urlo" di Munch



Fig. 5. Confronto tra le diverse texture ottenute enfatizzando lo stile del layer più esterno ($conv_1$) o quello più profondo ($conv_5$)

Come ultimo passo di questo lavoro, è stata analizzata l'evoluzione della funzione di loss in funzione delle epoche, considerando tutte le sue componenti. Si osserva che, nonostante il valore di β sia significativamente inferiore rispetto ad α , il contributo predominante alla loss proviene dallo stile. Questo risultato può essere spiegato dal fatto che l'immagine di partenza è identica a quella del contenuto, riducendo quindi al minimo la differenza tra le due immagini. Resta da stabilire dopo quante epoche l'immagine generata possa essere considerata accettabile, dal punto di vista della qualità del mix tra contenuto e stile.

IV. CONCLUSION

È stato dimostrato che, utilizzando una rete neurale convoluzionale pre-addestrata per il riconoscimento delle immagini, è possibile estrarre separatamente la componente di contenuto e la componente di stile di un'immagine. Inoltre, è possibile combinare la componente di stile di un'immagine con la componente di contenuto di un'altra, generando così una nuova immagine che incorpora entrambe le caratteristiche. Questa tecnica è nota come *Neural Style Transfer*.

Nell'ultima parte del lavoro, abbiamo analizzato

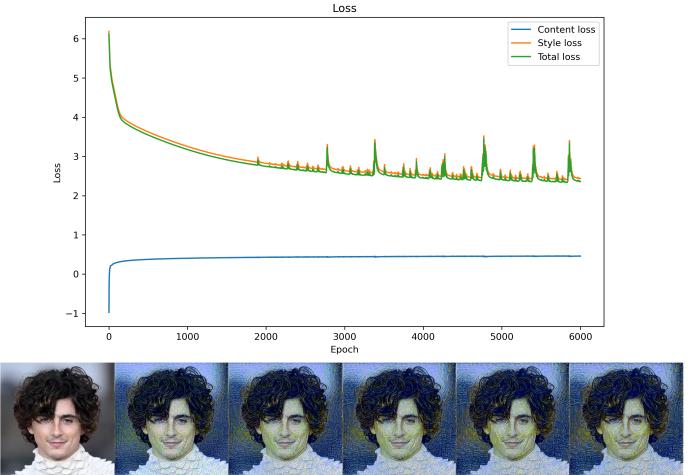


Fig. 6. Andamento della loss in funzione delle epoche. Sotto sono riportate le immagini ottenute ogni 1000 epoche

l'andamento delle diverse funzioni di loss. Ciò che è emerso chiaramente è la mancanza di una tecnica oggettiva per valutare la qualità delle immagini generate. Una possibile soluzione potrebbe essere quella di mostrare le immagini generate con diversi parametri e addestrate per un numero variabile di epoche a un campione significativo di persone, chiedendo loro di indicare quale immagine considerano migliore. I parametri potrebbero quindi essere ottimizzati sulla base delle statistiche raccolte.

Un'altra idea sarebbe quella di addestrare una CNN specificamente per il riconoscimento di opere di Van Gogh, quindi generare una versione *Notte stellata* di Timothée Chalamet e verificare, per quante epoche e per quali valori dei parametri, la rete riconosce l'immagine come appartenente allo stile di Van Gogh.

Infine, ulteriori implementazioni potrebbero concentrarsi sull'ottimizzazione del codice per ridurre i tempi di generazione delle immagini, migliorando così l'efficienza complessiva del processo.

REFERENCES

- [1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv:1508.06576 [cs.CV]*, 2015. [Online]. Available: <http://arxiv.org/abs/1508.06576>.
- [2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks," *arXiv:1505.07376 [cs, q-bio]*, 2015. [Online]. Available: <http://arxiv.org/abs/1505.07376>.
- [3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, 2014. [Online]. Available: <http://arxiv.org/abs/1409.1556>.