Progetto Machine Learning

AttnGAN e DualAttnGAN per Text-to-Image Synthesis

Giulia Giusti

Matteo Trentin

Dipartimento di Informatica
Università di Bologna

Text-to-Image Synthesis e obiettivi del progetto

- **Text-to-Image Synthesis**: traduzione di descrizioni testuali sotto forma di parole chiave o frasi in immagini con significato semantico simile al testo considerato.

- **Obiettivi del progetto**:
  - Studio alcune architetture di rete riguardanti l'approccio generativo al Text-to-Image Problem, nello specifico:
    - AttnGAN
    - DualAttnGAN
  - Riproduzione i risultati della rete AttnGAN
  - Implementazione della rete DualAttnGAN
  - Confronto dei risultati ottenuti da AttnGAN e DualAttnGAN

Strumenti e dataset

- **Dataset**: CUB contenente:
  - 11.788 immagini divise in 200 categorie
  - Annotazioni associate ad ogni immagine che ne descrivono le caratteristiche in forma testuale

- **Codice di base per il modello**: repository git relativo al codice di AttnGAN: https://github.com/taoxugit/AttnGAN

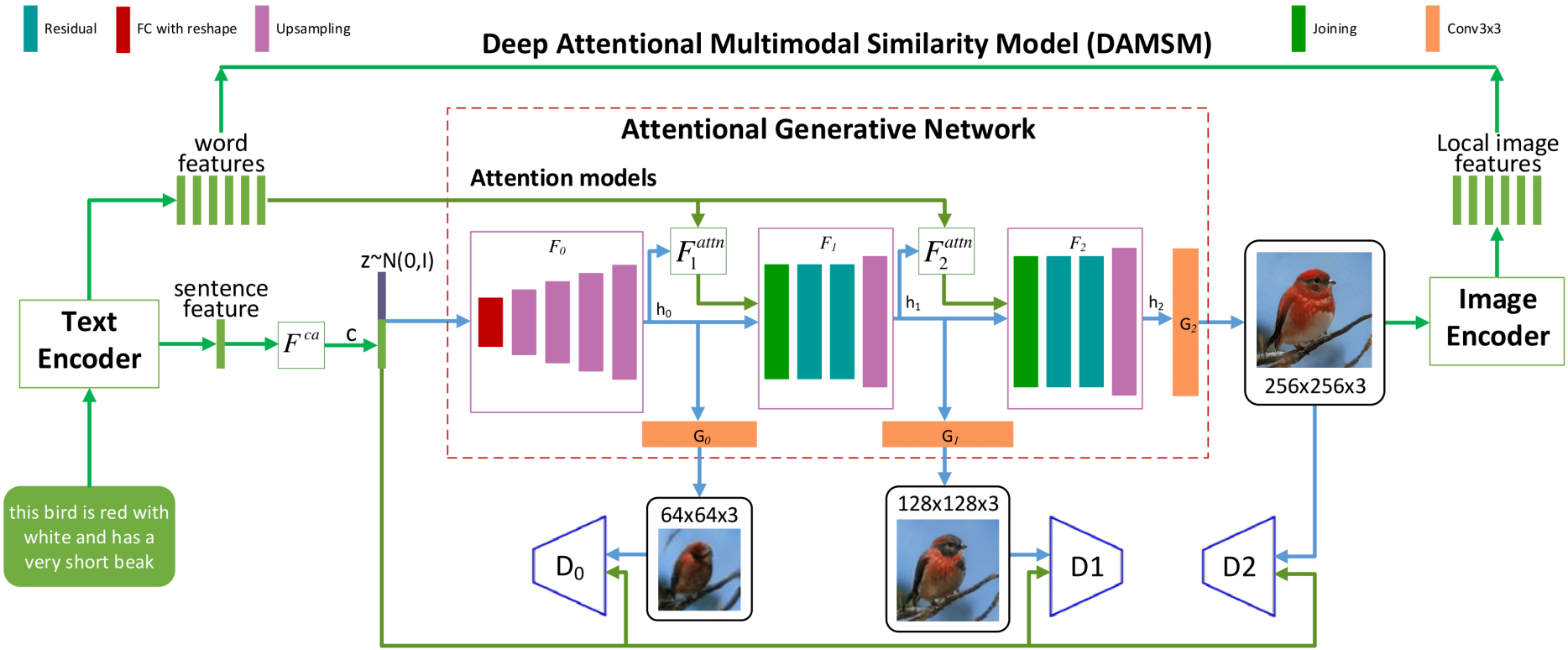
Struttura del progetto

1. Fase 1: Riproduzione dei risultati di AttnGAN
   - Rete AttnGAN
   - Riproduzione dei risultati di AttnGAN
2. Fase 2: Implementazione di DualAttnGAN
   - Rete DualAttnGAN
   - Modifiche al codice di AttnGAN
   - Esecuzione di DualAttnGAN e risultati ottenuti
3. Fase 3: Confronto AttnGAN e DualAttnGAN
   - Metriche usate per il confronto
   - Evaluation di AttnGAN
   - Evaluation di DualAttnGAN
   - Confronto tra AttnGAN e DualAttnGAN
4. Conclusioni e sviluppi futuri

Rete AttnGAN

La rete AttnGAN è formata da due componenti:

- Attentional Generative Network
- Deep Attentional Multimodal Similarity Model (DAMSM)



Rete AttnGAN - Attentional Generative Network 1/3

- **Obiettivo**: consentire alla rete di disegnare diverse sottoregioni dell'immagine in base alle parole della descrizione che sono più rilevanti.

- Essa è formata da m generator identificati da $G_0, ..., G_{m-1}$ che:
  - *Input*: hidden state della fase precedente calcolato dalla rete F_{i-1}, gli hidden state sono identificati da $h_0, ..., h_{m-1}$
  - *Output*: immagini di dimensioni crescenti all'aumentare di i $\widehat{x}_0, ..., \widehat{x}_{m-1}$.

$\Rightarrow \widehat{x}_i = G_i(h_i)$

Rete AttnGAN - Attentional Generative Network 2/3

Gli hidden state sono definiti come segue:

- $h_0 = F_0(z, F^{ca}(\bar{e}))$: nella prima fase viene utilizzato solo il vettore $\bar{e}$ che identifica la global sentence per generare un'immagine a bassa risoluzione.
  La rete F_0 è composta da: un layer lineare FC e 4 layer di upsampling.

- $h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1}))$ con $i > 0$: nelle fasi successive le reti utilizzano un vettore di contesto di parola che identifica le parole rilevanti per ogni sottoregione dell'immagine, tale vettore è viene ottenuto come output dell'attention model F_i^{attn}
  La rete F_i con $i > 0$ è composta da: un layer di joining, due layer residui e un layer di upsampling.

Rete AttnGAN - Attentional Generative Network 3/3

- **Obiettivo Attention Model**: identificare le parole rilevanti per la generazione di una determinata sottoregione dell'immagine.

- Input di F_i^{attn}:
  - *word features* rappresentate dal vettore $e \in \mathbb{R}^{DxT}$
  - *image features* rappresentate dall'hidden state $h \in \mathbb{R}^{\hat{D}xN}$

- Le word features e vengono convertite nel medesimo spazio semantico delle image features ottenendo $e' \in \mathbb{R}^{\hat{D}xT}$

- Calcolo del *word context vector* c_j per ogni sottoregione j dell'immagine, tale vettore contiene la rappresentazione delle parole rilevanti per la j-esima sottoregione dell'immagine.

- Output: $c = (c_0, ..., c_{N-1}) \in \mathbb{R}^{\hat{D}xN}$

Rete AttnGAN - DAMSM

- Il DAMSM esegue le seguenti operazioni:
  - Mappa le sottoregioni dell'immagine e le parole della frase in uno spazio semantico comune
  - Misura la somiglianza immagine-testo a livello di parola

- **Text Encoder**: LSTM che estrae due vettori semantici dal testo della descrizione presa in input:
  - word features $e \in \mathbb{R}^{DxT}$
  - sentence features $\bar{e} \in \mathbb{R}^{D}$

- **Image Encoder**: CNN che prende in input l'immagine generata dall'Attentional Generative Network, in cui i layer intermedi apprendono le features locali di diverse sottoregioni dell'immagine, mentre i layer successivi apprendono le caratteristiche globali dell'immagine.

Riproduzione dei risultati di AttnGAN

- Creazione di un ambiente isolato utilizzando `conda` contenente, tra le dipendenze:
  - `Pytorch 0.4.0`
  - `cuda92`
  - `cuDNN 7.1.2`
- Training e generazione delle immagini utilizzando gli iperparametri consigliati nell'articolo relativo ad AttnGAN
- Esecuzione di 600 epoch e generazione delle immagini ogni 50 epoch

Figura: Alcune immagini generate da AttnGAN dopo 600 epoch

Struttura del progetto

1. Fase 1: Riproduzione dei risultati di AttnGAN
   - Rete AttnGAN
   - Riproduzione dei risultati di AttnGAN
2. Fase 2: Implementazione di DualAttnGAN
   - Rete DualAttnGAN
   - Modifiche al codice di AttnGAN
   - Esecuzione di DualAttnGAN e risultati ottenuti
3. Fase 3: Confronto AttnGAN e DualAttnGAN
   - Metriche usate per il confronto
   - Evaluation di AttnGAN
   - Evaluation di DualAttnGAN
   - Confronto tra AttnGAN e DualAttnGAN
4. Conclusioni e sviluppi futuri

Rete DualAttnGAN-Differenze rispetto ad AttnGAN

- Lo stage iniziale ha la stessa struttura descritta per la rete iniziale F_0 di AttnGAN, con l'unica differenza che al posto dell'attivazione GLU viene utilizzata una ReLu.

- Negli stage successivi, il modulo di attention e il layer di joining vengono sostituiti dal Dual Attention Module (DAM) con l'obiettivo di migliorare i dettagli delle immagini generate dalla rete.

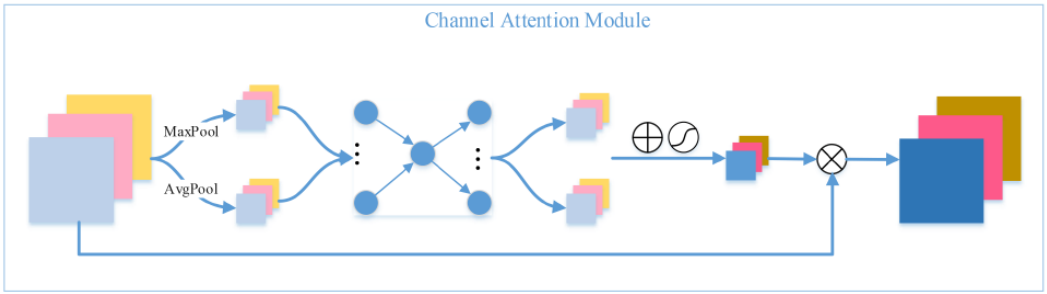
Rete DualAttnGAN-Dual Attention Module 1/3

Il Dual Attention Module della rete DualAttnGAN è composto dalle seguenti parti:

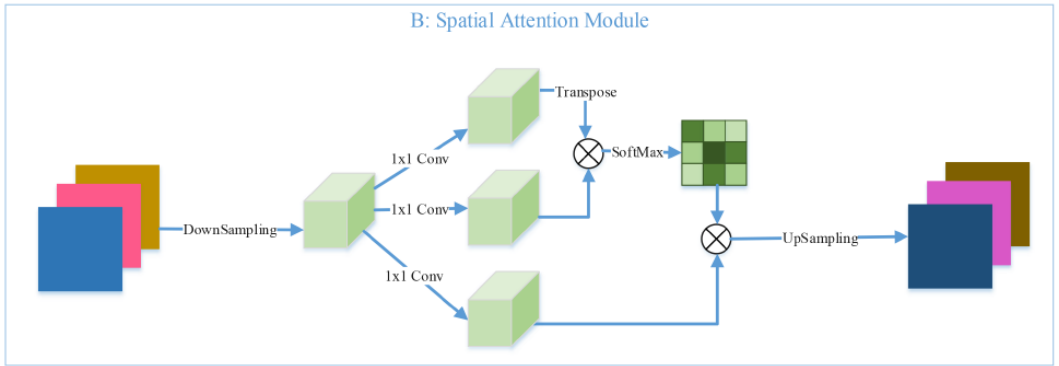
- Textual Attention Module (TAM) equivalente al modulo attention della rete AttnGAN, descritto nella funzione F^{attn}.
- Visual Attention Module (VAM): componente che modella le rappresentazioni interne dell'immagine rispetto al canale e agli assi spaziali, in modo tale da estrarre al meglio le strutture globali dell'immagine.
- Attenion Embedding Module (AEM) per unire e combinare le features estratte dall'analisi testuale del TAM e dall'analisi visuale del VAM.

Visual Attention Module (VAM)

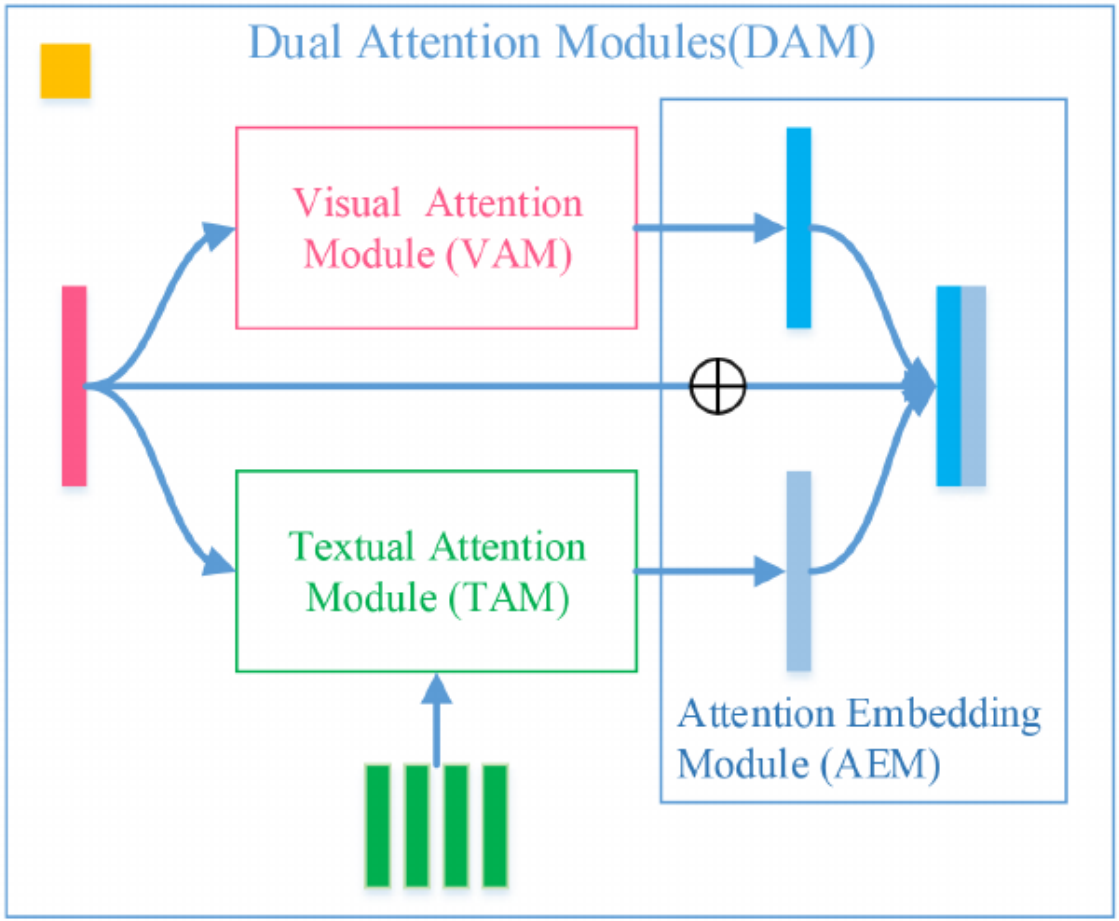
- Channel Attention Module:



- Spatial Attention Module:



Attenion Embedding Module (AEM)



Modifiche al codice di AttnGAN

- Modifiche presenti in `dualAttnModel.py`
- Implementati Channel Attention Module, Spatial Attention Module e Attention Embedding Module
- Implementata compressione tramite layer `conv2d` per migliorare l'utilizzo di memoria in Channel Attention Module

Esecuzione di DualAttnGAN e risultati ottenuti

Figura: Alcune immagini generate da DualAttnGAN dopo 250 epoch

Struttura del progetto

1. Fase 1: Riproduzione dei risultati di AttnGAN
   - Rete AttnGAN
   - Riproduzione dei risultati di AttnGAN
2. Fase 2: Implementazione di DualAttnGAN
   - Rete DualAttnGAN
   - Modifiche al codice di AttnGAN
   - Esecuzione di DualAttnGAN e risultati ottenuti
3. Fase 3: Confronto AttnGAN e DualAttnGAN
   - Metriche usate per il confronto
   - Evaluation di AttnGAN
   - Evaluation di DualAttnGAN
   - Confronto tra AttnGAN e DualAttnGAN
4. Conclusioni e sviluppi futuri

Metriche usate per il confronto

- **Inception Score (IS)** $\rightarrow$ Codice utilizzato
  - Applica una rete Inception-v3, allenata precedentemente su database ImageNet, alle immagini generate
  - Confronta la distribuzione condizionata e quella marginale per le categorie inferite dalla rete mediante Kullback Leibler Divergence

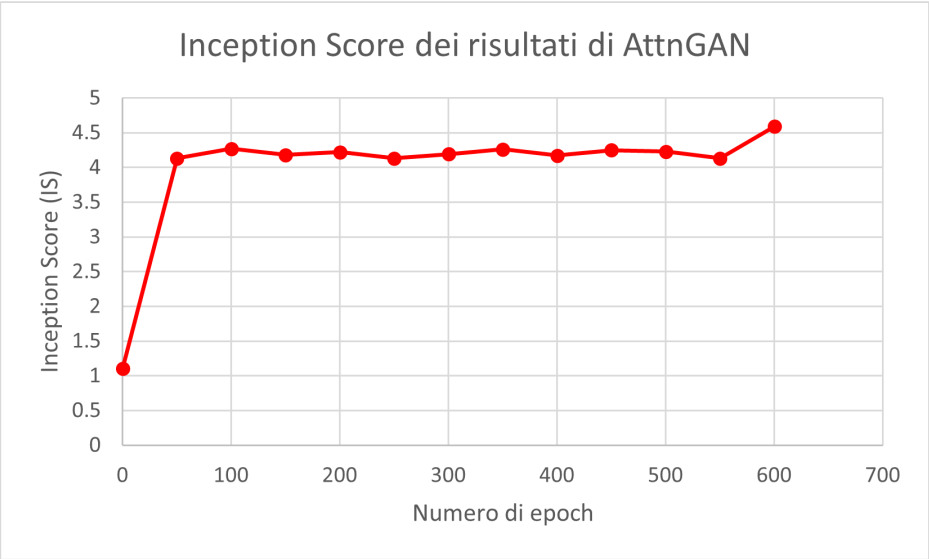
  $\Rightarrow$ IS più alto rappresenta immagini generate di migliore qualità

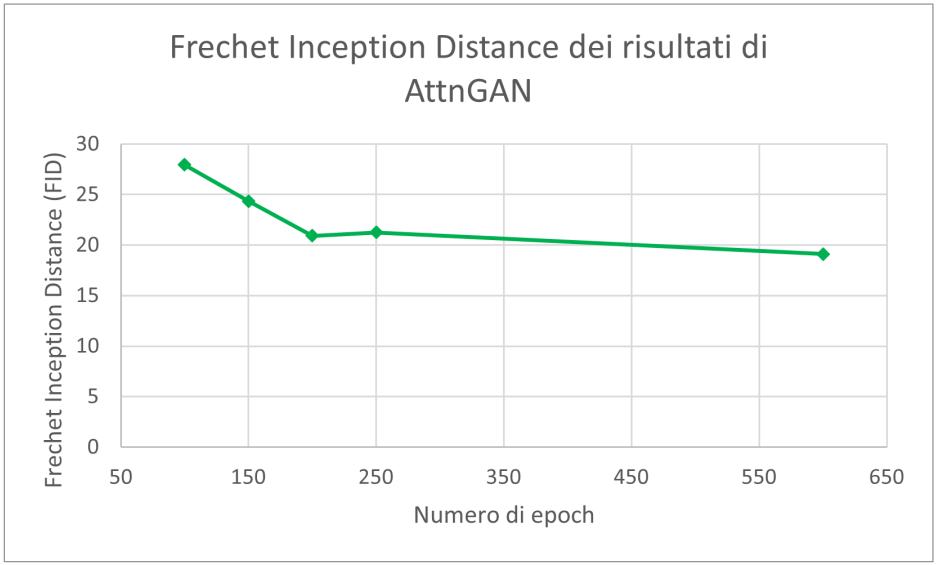
- **Frechet Inception Distance (FID)**:$\rightarrow$ Codice utilizzato
  Misura la distanza tra due curve:
  - La distribuzione delle futures delle immagini reali (immagini del dataset CUB)
  - La distribuzione delle futures delle immagini generate (immagini generate da AttnGAN o DualAttnGAN)

Evaluation di AttnGAN

In fase di evaluation di AttnGAN le metriche sono state calcolate sui risultati delle seguenti epoch:

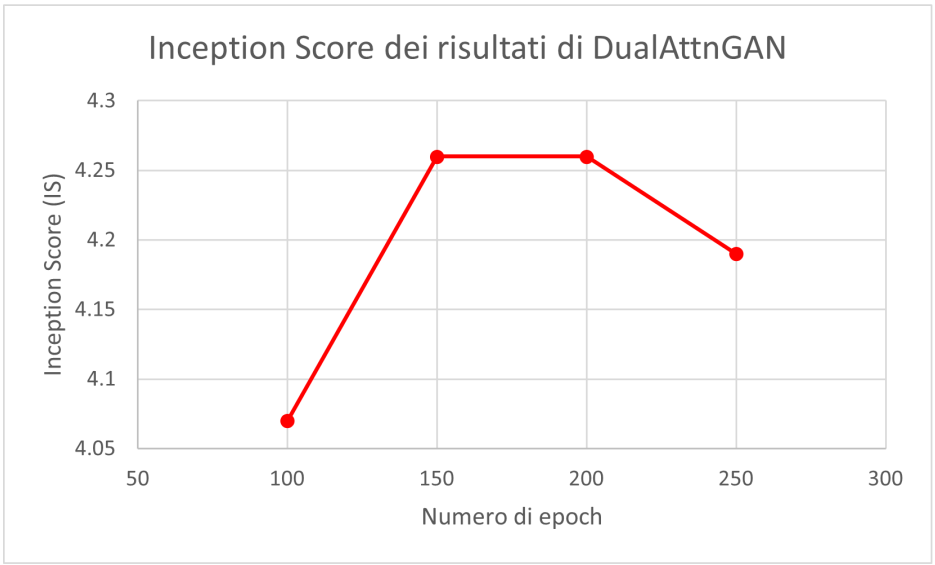
- Inception Score ogni 50 epoch
- FID per le epoch: 100, 150, 200, 250, 600

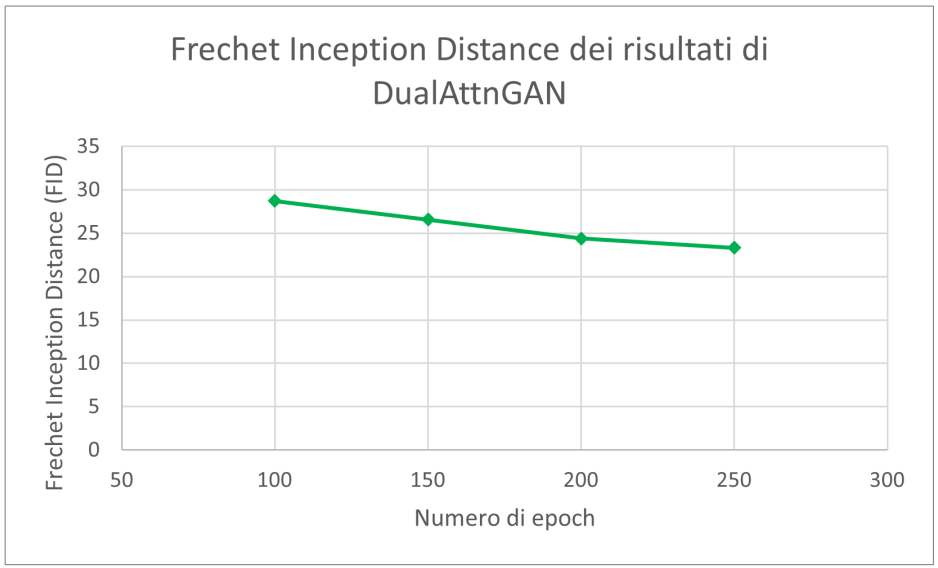




Evaluation di DualAttnGAN

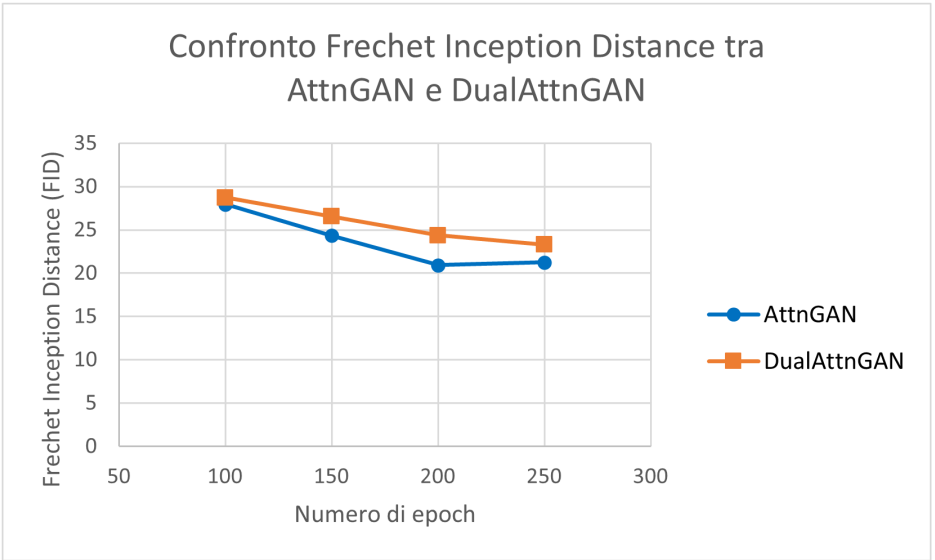
In fase di evaluation di DualAttnGAN entambe le metriche sono state calcolate sui risultati delle seguenti epoch: 100, 150, 200, 250.

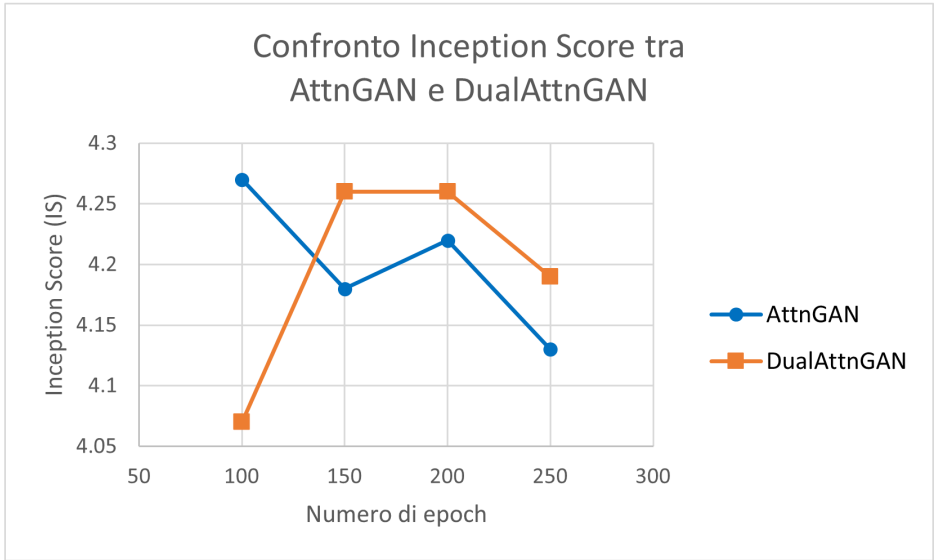




Confronto tra AttnGAN e DualAttnGAN

Confronto tra AttnGAN e DualAttnGAN su IS e FID calcolati ai riusati delle epoch: 100, 150, 200, 250.





Struttura del progetto

1. Fase 1: Riproduzione dei risultati di AttnGAN
   - Rete AttnGAN
   - Riproduzione dei risultati di AttnGAN
2. Fase 2: Implementazione di DualAttnGAN
   - Rete DualAttnGAN
   - Modifiche al codice di AttnGAN
   - Esecuzione di DualAttnGAN e risultati ottenuti
3. Fase 3: Confronto AttnGAN e DualAttnGAN
   - Metriche usate per il confronto
   - Evaluation di AttnGAN
   - Evaluation di DualAttnGAN
   - Confronto tra AttnGAN e DualAttnGAN
4. Conclusioni e sviluppi futuri

Conclusioni

- Non è stato possibile un confronto obiettivo tra le due architetture di rete a causa di un malfunzionamento hardware.
- La nuova implementazione risulta comparabile, in termini di prestazioni, con quella precedente, e mostra potenzialità di miglioramento su un training completo di 600 epoch.
- Non è stato notato un guadagno visivo sensibile in termini di distorsione dell'immagine, ma non è noto il margine di miglioramento ottenibile completando la calibrazione.

Sviluppi futuri 1/2

- Implementazione completa delle modifiche descritte nell'articolo, includendo quindi anche gli inverted residual layer
- Completamento del training con batch size aumentato, per migliorare la stabilità, e su 600 epoch, per effettuare un vero confronto col precedente metodo

Sviluppi futuri 2/2

- Ulteriori test del modello su diversi dataset, nello specifico COCO e CelebA; per quest'ultimo, implementare un sistema di traduzione da tag (presenti nel dataset) a caption (attese invece dall'encoder per il testo in DualAttnGAN)

- Analisi e confronto dei risultati delle due reti mediante una o più metriche in grado di misurare la corrispondenza tra il testo preso in input e l'immagine generata (metrica di R precision utilizzata in AttnGAN)

Bibliografia

- Jorge Agnese et al. A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis. In: (2019).
- Tao Xu et al. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In: (2018).
- Yali Cai et al. DualAttn-GAN: Text to Image Synthesis with Dual Attentional Generative Adversarial Network. In: (2019).
- Repository github per riproduzione dei risultati di AttnGAN: https://github.com/taoxugit/AttnGAN
- Repository github per calcolo dell'Inception score: https://github.com/hanzhanggit/StackGAN-inception-model
- Repository github per calcolo della Frechet Inception Distance: https://github.com/bioinf-jku/TTUR
- Repository github con la nostra implementazione di DualAttnGAN: https://github.com/mattrent/AttnGAN