AWOL per Generazione di Pixel Art:da Linguaggio a Pokemon tramite Sintesi Parametrica

July 13, 2025

Marina Valant

Abstract

Questo progetto estende il framework AWOL (Analysis WithOut synthesis using Language) dalla generazione di forme 3D alla generazione di pixel art, concentrandosi specificatamente su personaggi in stile Pokemon. Utilizzando gli embeddings CLIP e un modello di flusso Real-NVP, apprende una mappatura dalle descrizioni testuali a rappresentazioni parametriche che guidano un generatore di pixel art personalizzato. Questo approccio dimostra che il framework AWOL può generalizzare con successo oltre la geometria 3D per generare pixel art strutturata da descrizioni in linguaggio naturale, aprendo nuove possibilità per la generazione controllabile di contenuti nel dominio dell'arte digitale.

1. Introduction

Il framework AWOL originale (Zuffi & Black, 2024) ha dimostrato un successo notevole nella generazione di forme 3D apprendendo mappature dallo spazio latente CLIP a parametri di modelli parametrici, superando i tradizionali renderer differenziabili. Questo lavoro esplora la traduzione di questo paradigma alla generazione di pixel art, mirando specificamente ai personaggi in stile Pokemon.

La motivazione deriva dalla natura strutturata della pixel art, che condivide similitudini con i modelli 3D parametrici: entrambi i domini beneficiano del controllo esplicito dei parametri e possono produrre variazioni semanticamente significative attraverso l'interpolazione dei parametri. I personaggi Pokemon, con i loro pattern visivi distintivi e descrizioni testuali ben definite, forniscono un banco di prova ideale per questo approccio.

Email: Marina Valant <valant.2088938@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

1.1. Obiettivi

- Adattare il framework AWOL per la generazione di pixel art 2D
- Apprendere mappature significative dal linguaggio ai parametri della pixel art
- Dimostrare capacità di controllo semantico e interpolazione
- Valutare la qualità dei personaggi generati in stile Pokemon

2. Metodologia

2.1. Panoramica dell'architettura

Il sistema segue la struttura AWOL con tre componenti principali:

- ***Embedding del Testo***: Il modello CLIP genera embeddings a 512 dimensioni dalle descrizioni testuali
- ***Modello di Flusso***: Real-NVP con masking appreso mappa gli embeddings CLIP allo spazio dei parametri
- ***Generatore di Pixel Art***: Rete deconvoluzionale genera pixel art 64×64 dai parametri

2.2. Componenti del Modello

2.2.1. Generatore di Pixel Art

Il generatore utilizza un'architettura deconvoluzionale che trasforma un vettore di parametri a 128 dimensioni in immagini 64×64

2.2.2. Modello di Flusso Real-NVP

Il modello Real-NVP implementa trasformazioni invertibili con maschere fisse per mappare gli embeddings CLIP allo spazio dei parametri

2.3. Dataset e Preprocessing

Utilizziamo il dataset "Pokemon LLAVA Images and Text Descriptions" (Dat) da Kaggle, contenente:

- Immagini di Pokemon in formato bytes
- Descrizioni testuali dettagliate
- 833 campioni

Il preprocessing include:

- Ridimensionamento delle immagini a 64×64 con interpolazione nearest-neighbor
- Normalizzazione nel range [-1, 1]
- Generazione di embeddings CLIP per le descrizioni testuali

2.4. Funzione di Loss

La funzione di loss combina due componenti:

1. *Loss MSE*: Confronto pixel-wise tra immagini generate e reali
2. *Loss CLIP*: Allineamento semantico tra immagini generate e testo

```
loss = loss_mse * 0.1 + loss_clip * 1.5
```

Questa combinazione assicura sia la fedeltà visiva che la coerenza semantica.

3. Esperimenti e Valutazione

3.1. dettagli di training

Durante la ricerca degli iperparametri la dimensione del batch è stata variata tra 16 e 128, il numero di epoche tra 100 e 500 (per valori elevati di batch). Anche se al crescere della dimesione del batch e delle epoche migliora visivamente l'accordo tra vero e generato nel training, non generava miglioramenti significativi nei risultati complessivi.

Scelte finali:

- **Batch size**: 32
- **Epoche**: 100
- **Hardware**: GPU CUDA

3.2. post processing per pixel art

Per ottenere un'estetica pixel art più autentica, applica:

1. *Quantizzazione dei colori*: Riduzione a 16 livelli per canale
2. *Interpolazione nearest-neighbor*: Per bordi definiti
3. *Visualizzazione con effetto pixel*: Mantenimento della griglia pixel

4. Risultati

4.1. Qualità della Generazione

Il modello dimostra capacità di:

- Generare immagini coerenti con le descrizioni testuali
- Mantenere l'estetica pixel art caratteristica
- Produrre variazioni semanticamente significative

4.2. Analisi delle Loss

Durante il training si osserva:

- *Loss MSE*: Convergenza graduale verso valori bassi
- *Loss CLIP*: Miglioramento dell'allineamento semantico
- *Similarità coseno*: Incremento progressivo della coerenza testo-immagine

5. Limitazioni

5.1. Limitazioni *Attuali*

- *Risoluzione*: Limitata a 64×64 pixel
- *Diversità*: Dipendente dalla varietà del dataset
- *Controllo fine*: Controllo limitato su dettagli specifici

6. Conclusioni

Questo progetto dimostra con successo l'adattabilità del framework AWOL oltre il dominio 3D originale. La generazione di pixel art Pokemon-style da descrizioni testuali mostra che i principi di sintesi parametrica e mappatura da linguaggio naturale possono essere efficacemente trasferiti a nuovi domini creativi.

L'approccio apre interessanti possibilità per:

- Strumenti di creazione artistica assistita
- Generazione procedurale di contenuti per videogiochi
- Esplorazione di nuove modalità di controllo creativo

Il successo di questa trasposizione suggerisce che il framework AWOL ha potenziale per ulteriori applicazioni in domini dove il controllo parametrico strutturato è vantaggioso.

References

Pokemon llava images and text descriptions. URL https://huggingface.co/datasets/diffusers/pokemon-llava-captions.

Zuffi, S. and Black, M. J. Awol: Analysis without synthesis using language, 2024. URL https://arxiv.org/abs/2404.03042.