

Strategie di Green Fine-Tuning per la Predizione di Proprietà Molecolari

Bozza di Progetto

4 dicembre 2025

1 Introduzione e Contesto

La predizione delle proprietà molecolari è difficile per la scarsità di dati etichettati e la complessità delle rappresentazioni; per questo si usano modelli pre-addestrati seguiti da fine-tuning. L'aumento di complessità dei modelli comporta però maggiori emissioni di CO₂. L'idea è di esplorare un criterio di early-stopping di fine-tuning di questi modelli pre-addestrati basato sulle emissioni.

2 Metodologia: Criterio di Green Early Stopping adattivo

L'idea è di effettuare prima un classico finetuning, andando a misurare le metriche e le emissioni. Successivamente verrà applicato un criterio di early-stopping del fine-tuning basato sulle emissioni. Questo sicuramente porterà a un calo di performance, ma l'idea è analizzare il trade-off tra riduzione performance e riduzione emissioni per capire quale fine-tuning è il più ecologicamente sostenibile.

2.1 Misurazione delle Emissioni

Le emissioni di carbonio (E) sono calcolate in tempo reale utilizzando la libreria *CodeCarbon*, che stima il CO_2 equivalente ($CO_2\text{-eq}$) basandosi sul consumo energetico dell'hardware (GPU/CPU) e l'intensità di carbonio della rete elettrica locale.

2.2 Definizione dell'Adaptive Accuracy-Emission Ratio

Definiamo il rapporto di efficienza istantanea AER_norm all'epoca i come:

$$AER_norm(i) = \frac{\% \Delta \text{Performance}_i}{\% \Delta \text{Emission}_i} \quad (1)$$

Dove:

- $\% \Delta \text{Performance}_i$: è la variazione percentuale della metrica di validazione (es. guadagno in ROC-AUC o riduzione percentuale del RSE) rispetto all'epoca $i - 1$.
- $\% \Delta \text{Emission}_i$: è l'incremento percentuale delle emissioni cumulative rispetto all'epoca $i - 1$.

2.3 Criterio di Arresto Basato su Media Mobile

Invece di confrontare AER_norm con una soglia statica, lo confrontiamo con la sua stessa storia, rappresentata da una Media Mobile Esponenziale (EMA). Questo permette all'algoritmo di "imparare" la velocità di convergenza specifica del modello (sia esso un Transformer pesante o una GNN leggera).

Algorithm 1 AN-GES semplificato

```
1: Input:  $\alpha$  (smoothing, es. 0.9),  $\beta$  (soglia, es. 0.2),  $W$  (warm-up epoche)
2:  $EMA_{AER} \leftarrow$  valore iniziale piccolo
3: for epoca  $i = 1, 2, \dots$  do
4:   calcola metrica di validazione e emissioni cumulative  $E_i$ 
5:   if  $i > 1$  then
6:     calcola  $\% \Delta \text{Perf}_i$  e  $\% \Delta \text{Emiss}_i$ 
7:      $AER_{current} \leftarrow \frac{\% \Delta \text{Perf}_i}{\% \Delta \text{Emiss}_i}$ 
8:     if  $i \geq W$  then
9:        $EMA_{AER} \leftarrow \alpha \cdot AER_{current} + (1 - \alpha) \cdot EMA_{AER}$ 
10:      if  $AER_{current} < \beta \cdot EMA_{AER}$  then
11:        Stop Training
12:      end if
13:    else
14:      accumula  $AER_{current}$  per inizializzare  $EMA_{AER}$ 
15:    end if
16:  end if
17: end for
```

3 Disegno Sperimentale

3.1 Dataset

Utilizzeremo i dataset della suite *MoleculeNet*

3.2 Modelli Oggetto di Studio

Confronteremo l'efficacia del criterio AN-GES su due famiglie di modelli pre-addestrati, caratterizzati da profili energetici molto diversi:

1. **Modelli 1D (Sequence-based):** Utilizzano stringhe SMILES/SELFIES.
2. **Modelli 2D (Graph-based):** Utilizzano la topologia del grafo molecolare.