Analisi e confronto del Vision Transformer con ResNet18 su ImageNet

Arianna Cipolla

10/06/2025

arianna.cipolla@studenti.unipr.it

UNIVERSITÀ
DI PARMA

Obiettivo

Confrontare le architetture ViT e ResNet18 in termini di prestazioni su ImageNet e valutarne la robustezza contro attacchi FGSM.

UNIVERSITÀ DI PARMA

Obiettivo

Confrontare le architetture ViT e ResNet18 in termini di prestazioni su ImageNet e valutarne la robustezza contro attacchi FGSM.

1 **Spiegazione** struttura reti neurali

2 **Confronto prestazioni** ViT VS ResNet18

3 **Attacco della rete** tramite FGSM

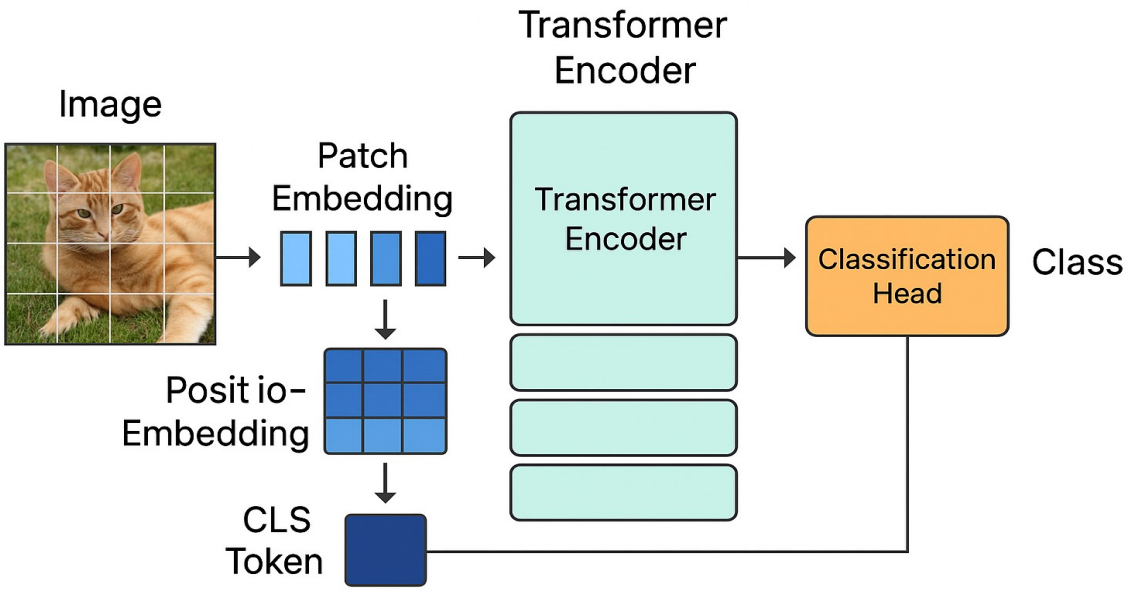
UNIVERSITÀ DI PARMA

Dataset utilizzato

ImmageNet mini con subset divisi in:

- Training set da 10, 50, 100, 150 e 200 cartelle con al suo interno circa 20 immagini
- Validation set composto nello stesso modo del Training set

Vision Trasformer



Vision Trasformer - implementazione

- Ho utilizzato la libreria **timm** che fornisce modelli pre-addestrati su ImageNet
  - Scegliendo la variante che lavora con immagini 224x224 pixel e suddivise in patch da 16x16 pixel
- Sostituisco la testa finale con un **layer fully connected** di dimensione pari al numero di classi da classificare
- Congelo i pesi di tutti gli strati tranne quelli della testa che l'ho allenata con l'ottimizzazione **Adam** e la **Cross-Empty Loss**

ResNet18

Un tipo di rete neurale convoluzionale composta da 18 layers che ha la proprietà di utilizzare i **blocchi residui:**

- Invece di passare solo attraverso i filtri convoluzionali, ogni blocco somma l'input iniziale al risultato dell'elaborazione, risolvendo il problema del **vanishing gradient**

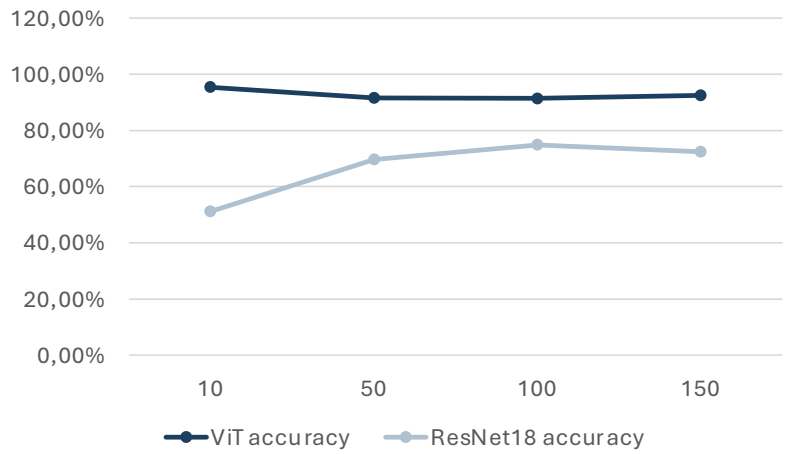
Implementazione:

Come nella ViT **la testa** è un layer fully connected che viene adattato al numero di classi che sto elaborando e congelo i pesi della parte precedente della rete

UNIVERSITÀ DI PARMA

Confronto risultati

Classi	ViT accuracy	ResNet18 Accuracy
10	**95.35%**	51.16%
50	91.67%	69.61%
100	91.44%	**74.82%**
150	92.47%	72.41%
200	91.12%	70.12%



UNIVERSITÀ
DI PARMA

Attacco delle reti con FGSM

Applico piccole **perturbazioni** di diversa potenza alle immagini per vedere se i modelli sono robusti

Originale

FGSM ε=0.01

UNIVERSITÀ DI PARMA

FGSM - risultati

Classi	Epsilon	ViT accuracy	ResNet18 Accuracy
10	0,01	**88.37%**	25.58%
	0,05	69.77%	23.26%
	0,1	60.47%	23.26%
50	0,01	85.29%	27.94%
	0,05	70.10%	18.14%
	0,1	57.84%	16.18%
100	0,01	85.82%	**29.34%**
	0,05	73.11%	15.65%
	0,1	61.86%	11.74%

Classi	Epsilon	ViT accuracy	ResNet18 Accuracy
150	0,01	83.44%	22.24%
	0,05	69.23%	14.05%
	0,1	59.20%	9.20%
200	0,01	84.39%	21.80%
	0,05	70.39%	13.86%
	0,1	60.30%	10.36%

UNIVERSITÀ DI PARMA

Conclusioni

Vision Transformer

- Prestazioni migliori su dataset grandi
- Ha una visione più globale dell'immagine quindi si difende meglio verso attacchi FGSM

ResNet18

- Più robusto su dataset piccoli
- Dipende dalle caratteristiche locali dei pixel, quindi più sensibile ad attacchi FGSM

UNIVERSITÀ DI PARMA

Bibliografia

- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- ResNet18
- Adversarial Example Generation
- Pytorch Image Models (timm)