ADVERSARIAL REPROGRAMMING DI MODELLI PRE-ADDESTRATI PER LA CLASSIFICAZIONE AUDIO: UN APPROCCIO DUAL-DOMAIN

Carlo Aironi, Leonardo Gabrielli, Emanuele Principi, Stefano Squartini

Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche
Via Brecce Bianche 12, 60131, Ancona, Italia, e-mail s.squartini@univpm.it

L'addestramento di modelli Deep Learning ad alte prestazioni richiede ingenti risorse computazionali, dataset di grandi dimensioni e un accurato tuning dell'architettura e degli iperparametri. In contesti con vincoli di memoria, tempo o scarsa disponibilità di dati, è cruciale valutare se modelli pre-addestrati in domini ricchi di risorse possano essere riutilizzati in domini differenti. L'Adversarial Reprogramming [1-3] affronta questo problema consentendo di sfruttare il bias induttivo di un modello pre-addestrato, senza modificarne i parametri di memoria, per un nuovo task, attraverso la sola introduzione di una trasformazione sull'input e, opzionalmente, una mappatura dell'output. Tale approccio permette il riutilizzo efficiente di modelli già ottimizzati in ambiti come visione artificiale o linguaggio naturale, minimizzando i parametri da apprendere e il costo computazionale.

Questo lavoro propone una metodologia di Adversarial Reprogramming per la classificazione audio, riutilizzando modelli pre-addestrati su task di visione artificiale (ad es. ImageNet). In particolare, viene presentata una strategia innovativa che applica trasformazioni in due differenti domini: temporale (waveform) e tempo-frequenza (spettrogramma). L'obiettivo è "riprogrammare" un modello convoluzionale di grandi dimensioni (*source model*) per svolgere un compito di Keyword Spotting (*target task*), con una quantità ridotta di dati e parametri.
Per mappare le predizioni dalle N_s classi source alle N_t classi target (tipicamente $N_s > N_t$), in letteratura sono state proposte varie strategie. Partendo dall'analisi di quest'ultime, nel lavoro vengono investigati due criteri: una mappatura statica basata su clustering nello spazio latente del modello source, e una mappatura parametrizzata. Quest'ultima, sebbene richieda complessivamente più memoria, ha mostrato prestazioni superiori in termini di accuratezza.

La *Figura 1* illustra il framework generale di Adversarial Reprogramming (a), introdotto in origine da Elsayed et al. [1], insieme al metodo dual-domain proposto per la classificazione di sequenze audio (b). Le parti evidenziate in giallo, a sinistra del *source model*, rappresentano i *Programs*, ovvero le perturbazioni additive che vengono sovrapposte all'input temporale (P_{1D}) e utilizzate come padding dell'input tempo-frequenza (P_{2D}), per guidare il modello verso l'output desiderato. La funzione di costo ottimizza congiuntamente i parametri dei *Programs* 1D e 2D e del layer di mappatura. L'ottimizzazione avviene minimizzando la cross-entropia tra predizioni mappate e target reali, includendo un termine di regolarizzazione per prevenire l'overfitting.

Gli esperimenti sono stati condotti utilizzando il dataset *Google Speech Commands* [4], limitatamente a dieci categorie di parole. Le architetture considerate come modello source sono *SqueezeNet v.1.1*, *MobileNet v.3* e *ResNet 50*, pre-addestrate su *ImageNet* (1000 classi). Le condizioni di addestramento includono sia l'uso dell'intero dataset (circa 31000 campioni), sia scenari di data-scarcity, con 100, 200, 500 e 1000 campioni totali (corrispondenti, rispettivamente a 10, 20, 50 e 100 campioni per classe), che rendono l'addestramento del modello particolarmente impegnativo. Inoltre, come metodo di confronto è stato considerato un approccio

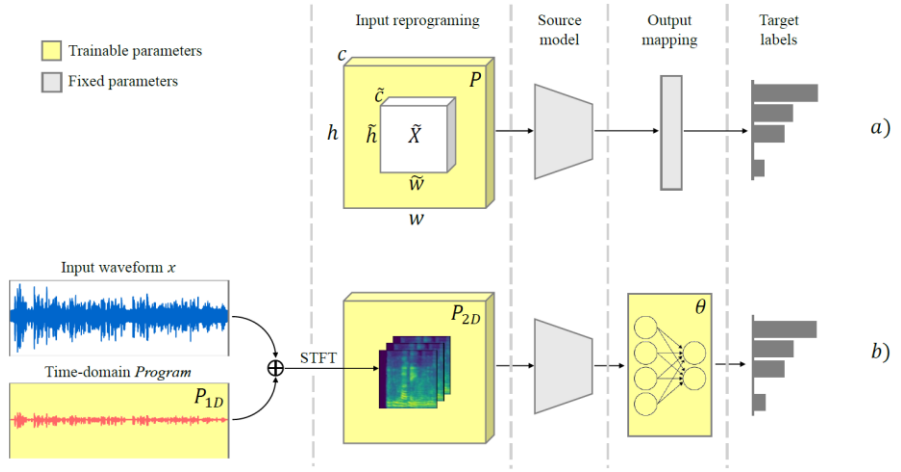

Figura 1 – Rappresentazione schematica degli approcci Adversarial Reprogramming, classico (a) e dual-domain (b).

tradizionale, basato su Depthwise-Separable Convolutional Neural Network (DSCNN), addestrato da zero (cioè senza Reprogramming) al task di classificazione di spettrogrammi. I risultati, riportati in *Tabella 1*, evidenziano che, in scenari a bassa disponibilità di dati, il framework proposto supera il metodo baseline DSCNN, fatta eccezione per il reprogramming della rete *SqueezeNet*, a causa della sua eccessiva compressione. La colonna di destra riporta il tempo necessario per un ciclo forward-backpropagation-update, riferito ad un singolo campione. Tali valori evidenziano un altro punto di forza del paradigma Reprogramming, diretta conseguenza del numero significativamente contenuto di parametri allenabili.

Tabella 1 – Accuratezza di classificazione e tempo stimato di addestramento per singolo campione.

	training samples					Training time
Model	100	200	500	1000	31000	(ms)
SqueezeNet v1.1	23.03%	38.75%	59.42%	61.69%	84.34%	20.50
MobileNet v3	47.61%	57.91%	63.84%	71.93%	89.55%	22.34
ResNet 50	**50.79%**	**58.20%**	**64.90%**	**77.59%**	92.10%	41.00
DSCNN	29.92%	44.02%	60.17%	73.06%	**95.03%**	96.43

Il lavoro dimostra dunque l'efficacia dell'Adversarial Reprogramming nel riutilizzo efficiente di modelli vision in contesti audio, rappresentando una soluzione promettente per l'apprendimento low-resource e su dispositivi a prestazioni limitate.

Bibliografia:

[1] – Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J.: Adversarial Reprogramming of neural networks. In: International Conference on Learning Representations, ICLR (2019).

[2] – Englert, M., Lazic, R.: Adversarial Reprogramming revisited. Advances in Neural Information Processing Systems 35, 28588–28600 (2022)

[3] – Zheng, Y., Feng, X., Xia, Z., Jiang, X., Demontis, A., Pintor, M., Biggio, B., Roli, F.: Why Adversarial Reprogramming works, when it fails, and how to tell the difference. Information Sciences 632, 130–143 (2023)

[4] – Warden, P.: Speech Commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)