Intelligent embedded systems for facial soft biometrics in social robotics

(sistemi integrati intelligenti per biometria “soft” nella robotica sociale)

Abstract – Vincenzo Vigilante

l fondamento di questa tesi è l'osservazione dell'utilità degli indizi contestuali nel contesto della robotica sociale: studi dimostrano che il comportamento simile all'uomo è la chiave per generare nell'interlocutore il sentimento di empatia che gli consente di percepire inconsciamente il robot come proprio pari .

Dall'analisi dei volti delle persone intorno, il robot può raccogliere informazioni che consentono di personalizzare l'interazione e potenziare la sensazione di empatia. Tali informazioni, tra cui età, sesso, etnia, emozione e altro, sono chiamate biometria “soft" perché non consentono l'identificazione univoca, perfetta, di una persona, ma sono comunque utilizzate dagli esseri umani per distinguere i loro coetanei;

Osserviamo che i compiti nel dominio della biometria “soft” facciale sono ampiamente studiati in letteratura, ma l'applicazione a condizioni realistiche come quelle della robotica sociale, introduce alcuni vincoli che richiedono un'attenzione specifica, vale a dire vincoli di risorse e vincoli di robustezza. I vincoli di risorse sono limitazioni dovute all'hardware effettivo che esegue i sistemi di previsione; tali vincoli, ad esempio, richiedono che l'impronta della memoria sia limitata a ciò che l'hardware può gestire e richiedono che anche il tempo di inferenza sia limitato, in modo che le informazioni siano disponibili in tempo utile per essere utilizzate in un'iterazione a ritmo naturale. La robustezza riguarda la capacità del sistema di produrre previsioni corrette basate su immagini di input che sono influenzate da tutti i tipi di corruzioni e perturbazioni che sono presenti su immagini acquisite in condizioni non vincolate utilizzando hardware tipico dell'applicazione considerata; ad esempio, le telecamere incorporate nei robot producono immagini rumorose con una risoluzione e una gamma dinamica limitate.

In questa tesi affrontiamo questi temi nel contesto del Deep Learning. Progettiamo e valutiamo metodi basati sulla CNN efficienti ed efficaci per le attività di riconoscimento del genere, riconoscimento dell'etnia, stima dell'età e classificazione delle emozioni.

Per il riconoscimento del genere osserviamo che le architetture CNN tradizionali sono progettate con riferimento al problema del riconoscimento degli oggetti e valutate sul benchmark ImageNet. Sosteniamo che il compito del riconoscimento del genere abbia caratteristiche diverse, quindi progettiamo un'architettura efficiente basata sull'architettura MobileNet v2 ma con profondità, dimensioni di input e numero di mappe di funzionalità ridotte. Sperimentiamo la nostra architettura ridotta sul benchmark pubblico LFW + e la confrontiamo con lo stato dell'arte. Dimostriamo che la nostra architettura è in grado di riconoscere il genere con una precisione del 98,1% in soli 56 ms su un dispositivo “embedded” senza alcuna accelerazione della rete neurale, che riteniamo perfettamente ragionevole per l'applicazione ai robot sociali.

Quando si considera lo stato dell'arte del riconoscimento dell'etnia, si conclude che lo sviluppo di metodologie efficaci è frenato dall'assenza di un ampio dataset. Progettiamo efficacemente un dataset di grandi dimensioni (3,3 milioni di immagini, 9000 identità diverse) e annotato in modo affidabile, facendo partecipare più persone di etnie diverse all'annotazione degli stessi dati. Sosteniamo che l'addestramento sul nostro dataset renda le reti neurali più accurate rispetto all'addestramento su altri dataset; lo dimostriamo addestrando diverse reti neurali, inclusa l'architettura efficiente MobileNet v2 citata in precedenza, e utilizzando un benchmark indipendente per valutare che la precisione è effettivamente maggiore. Riteniamo che i nostri risultati abbiano grandi margini di miglioramento, infatti li consideriamo come una baseline e rendiamo il nostro dataset (VMER) pubblicamente disponibile per promuovere lo sviluppo dello stato dell'arte su questo compito.

Per quanto riguarda il riconoscimento dell'età, quando si cerca di addestrare un'architettura veloce per questo compito (come abbiamo fatto per il genere e l'etnia) troviamo che i dataset disponibili pubblicamente

sono troppo piccoli o contengono un'annotazione automatica molto rumorosa; questo è comprensibile, poiché annotare un dataset contenente milioni di immagini è un'attività estremamente costosa. I metodi accurati esistenti che sono stati proposti fino ad ora, ricorrono all'uso di costose procedure di pulizia manuale e utilizzano grandi ensemble di reti neurali che sono più resistenti all'inadeguatezza del dataset rispetto alle architetture più semplici. Questi ensemble sono ovviamente lenti e grandi e inadeguati per applicazioni pratiche; per esempio, il vincitore della prestigiosa sfida LAP 2016 impiega circa 6 secondi per immagine e richiede l'utilizzo di una potente GPU: visti i vincoli della nostra applicazione, una tale soluzione non è fattibile per noi.

Proponiamo l'uso di una tecnica chiamata distillazione della conoscenza (knowledge distillation) per superare il problema: usiamo l'ensemble potente ma lento che abbiamo appena descritto (che chiamiamo insegnante) per annotare un dataset molto grande (che chiamiamo VMAGE), e poi lo usiamo per addestrare architetture semplici (studenti). Dimostriamo che i nostri studenti superano consistentemente le stesse architetture addestrate con il dataset comunemente usato dallo stato dell'arte, senza l'uso della tecnica di distillazione. Ciò dimostra l'efficacia del nostro approccio. Inoltre, confrontiamo l'accuratezza delle reti addestrate con il nostro metodo su tutti i principali benchmark pubblici e dimostriamo che, con diversi protocolli, esse raggiungono un'accuratezza competitiva rispetto alla letteratura esistente, pur essendo più semplici ed efficienti, quindi adeguate per l'uso nella nostra applicazione proposta.

Infine, con riferimento al compito di Emotion Recognition, sperimentiamo l'effetto di differenti corruzioni e perturbazioni dell'immagine dal mondo reale su 4 differenti architetture (VGG, SENet, DenseNet e Xception). Valutiamo l'effetto dell'Autoaugment e del downsampling con antialias su tali architetture, la prima è una tecnica per aumentare efficacemente i dati di addestramento e la seconda è una modifica dell'architettura che aggiunge filtri passa-basso all'interno delle reti ovunque si verifichi un downsampling (es. Max-pooling o strided convoluzioni). Per la nostra valutazione, costruiamo un dataset di benchmark in cima al set di test RAF-DB che include immagini con corruzioni che tipicamente si verificano quando i sistemi di riconoscimento vengono distribuiti in scenari reali. Le corruzioni includono diversi tipi di sfocatura (sfocatura movimento, sfocatura lente, sfocatura zoom, sfocatura gaussiana), di rumore (rumore gaussiano, rumore di ripresa), pixelazione, compressione jpeg, cambiamenti di luminosità e contrasto e combinazione di questi. Per valutare la stabilità delle previsioni, generiamo il dataset RAF-DB-P, che include versioni delle immagini di prova in cui perturbiamo la luminosità, la posizione, la scala, la rotazione, la quantità di sfocatura e il pattern del rumore. Troviamo che l'uso combinato di antialiasing e Autoaugment contribuisce in modo sostanziale al miglioramento della robustezza alle corruzioni, specialmente a quelle di tipo noise e digitale, di SENet e DenseNet. L'architettura VGG ha invece mostrato la massima stabilità di classificazione rispetto alle perturbazioni che interessano i successivi frame di una sequenza, soprattutto se abbinata all'utilizzo di filtri anti-aliasing. Riteniamo che i metodi Xception non siano adatti per l'analisi delle emozioni facciali nel nostro ambiente, poiché sono particolarmente influenzati da corruzioni e perturbazioni. In conclusione, i nostri esperimenti hanno dimostrato che le corruzioni e le perturbazioni comuni sono aspetti importanti da tenere in considerazione quando si valutano metodi da utilizzare in scenari reali. Tuttavia, nessuno dei metodi esistenti, che abbiamo modificato con filtri anti-aliasing e addestrati con un’ampia augmentation dei dati, ha mostrato robustezza a tutte le corruzioni e perturbazioni considerate, quindi questo aspetto rimane aperto per indagini future.

Complessivamente, in questo lavoro, siamo stati in grado, per ciascuno dei quattro compiti, sesso, età, etnia, emozione, di progettare un sistema basato sulla CNN in grado di raggiungere prestazioni allo stato dell'arte pur essendo in grado di esibirsi nell'ambiente robotico sociale di destinazione, con tempi di inferenza e requisiti di memoria limitati e in grado di lavorare in contesti ragionevolmente "wild", incontrollati. Il lavoro futuro dovrà concentrarsi sul problema della robustezza e proporre una soluzione per un sistema di percezione più affidabile.