Università degli Studi di Salerno

.DIEM

Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata

Dottorato di Ricerca in Ingegneria dell'Informazione
Ciclo 35

TESI DI DOTTORATO / PH.D. THESIS

Social Robots, from intelligent perception to emphatic behaviors

ANTONIO ROBERTO

SUPERVISOR: **PROF. MARIO VENTO**
PROF. LUC BRUN

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Anno 2023

Abstract

I robot sociali sono una particolare categoria di robot in grado di percepire informazioni sull'ambiente circostante, ragionare su tali informazioni e interagire con gli esseri umani. Tra le notevoli applicazioni dei robot sociali vi sono le guide museali, gli assistenti ospedalieri, il trattamento dell'autismo, gli assistenti alberghieri e l'assistenza agli anziani. Gli studi dimostrano che la capacità dei robot sociali di personalizzare la conversazione e percepire le emozioni dell'interlocutore sono i comportamenti chiave che permettono loro di essere considerati empatici. A tale scopo, i robot sociali si sfruttano diverse modalità sensoriali per acquisire informazioni sull'interlocutore in modo robusto e preciso. L'equipaggiamento sensoriale è cruciale considerando che questo tipo di robot lavora comunemente in ambienti non controllati, ad esempio in condizioni di luce dinamiche e con forte rumore ambientale. Oltre a queste sfide, i robot sociali devono conversare in tempo reale con gli esseri umani per dare loro la sensazione di interazione naturale. Considerando il contesto applicativo, questo risultato è possibile solo se gli algoritmi vengono eseguiti a bordo della piattaforma del robot sociale, il che rende il compito più difficile a causa dei vincoli impliciti di calcolo e di potenza. Questa tesi affronta queste esigenze nel contesto del Deep Learning. In particolare, è stata proposta una nuova architettura software ottimizzata per interazioni multimodali in tempo reale come soluzione generale per i robot

sociali. La realizzazione di un prototipo robotico ha permesso di identificare i principali problemi percepiti dagli esseri umani riguardo agli algoritmi all'avanguardia relativi all'interazione uomo-robot quando utilizzati insieme in un'applicazione reale. Alla luce di questo risultato, questa tesi avanza lo stato dell'arte proponendo e validando nuovi algoritmi di comprensione del linguaggio naturale e dell'analisi audio ottimizzati per essere eseguiti su sistemi embedded robotici, mantenendo un'alta accuratezza. L'architettura del robot sociale proposta include tutti i moduli software che consentono di soddisfare i principali requisiti di un robot sociale: primo, un pianificatore di dialogo in grado di personalizzare l'interazione uomo-robot sfruttando i dati biometrici rilevati dai sensori del robot sociale; secondo, un modulo di aggregazione di sensori multimodali in grado di sfruttare le informazioni acquisite da diversi tipi di sensori per aumentare la robustezza al rumore ambientale; infine, delle pipeline di elaborazione parallele che, progettate e implementate correttamente, garantiscono prestazioni in tempo reale. È stato realizzato un prototipo di robot sociale basato sull'architettura proposta, che è stato utilizzato durante l'esposizione SICUREZZA per tre giorni. 161 persone hanno interagito con il robot e hanno valutato la loro esperienza rispondendo a 5 domande con un punteggio tra 1 e 5. Il punteggio massimo è stato raggiunto per più del 40% delle risposte e il tasso medio si situava tra 4 e 5. Questo risultato acquista maggior rilevanza considerando che le persone che hanno partecipato alla conferenza erano tecnicamente preparate e, quindi, il loro giudizio è affidabile. L'indagine ha anche permesso di esplorare le sensazioni degli umani sulla performance degli algoritmi all'avanguardia disponibili sul prototipo proposto. Questa analisi ha evidenziato la necessità di algoritmi audio più robusti al rumore ambientale e di pipeline di elaborazione del dialogo più efficienti. Partendo da queste evidenze, questa tesi propone inizialmente due rappresentazioni audio apprendibili per raggiungere i seguenti obiettivi: robustezza al rumore ambientale ed efficienza computazionale. In

primo luogo, è stato proposto un nuovo livello convoluzionale, chiamato Denoising-Enhancement Layer (DELayer). È in grado di pulire dal rumore e migliorare il segnale in ingresso combinando un modulo di attenzione (DELayer) con il livello SincNet; la rappresentazione finale è quindi in grado di attenuare le componenti di frequenza affette dal rumore ambientale e amplificare quelle rilevanti per la classificazione. I risultati sperimentali dimostrano che il modello end-to-end basato su DELayer, ovvero DENet, è decisamente più efficace dei metodi esistenti nel rilevare e riconoscere eventi audio di interesse sui benchmark MIVIA Audio Events e MIVIA Road Events. La nuova CNN raggiunge prestazioni all'avanguardia su entrambi i dataset e ulteriori analisi mostrano la sua robustezza al rumore, la sua stabilità tra diverse condizioni ambientali e le sue capacità di generalizzazione. La valutazione della latenza del modello DENet ha evidenziato la necessità di un modello audio ottimizzato per essere eseguito su dispositivi CPU senza "batching" delle previsioni così da aumentare la velocità del modello stesso. Per questo motivo, in questa tesi i fondamenti di DELayer sono stati estesi a DEGram, una rappresentazione apprendibile basata su spettrogramma. Essendo una rappresentazione simile a uno spettrogramma, DEGram consente l'uso di CNN più piccole con un'accuratezza comparabile. Di conseguenza, riduce i requisiti computazionali del sistema finale mantenendo la robustezza di DENet. DEGram è il risultato della combinazione di due nuovi strati: SincGram e una versione tempo-frequenza di DELayer, chiamata TF-DELayer. Il primo, come SincNet, è in grado di apprendere le frequenze di interesse per il problema di analisi audio che stiamo affrontando, estrarre le feature specifiche del problema attraverso filtri passa-banda addestrabili. Il secondo può rimuovere il rumore di fondo ambientale dall'input audio utilizzando un meccanismo di attenzione tempo-frequenza, concentrando la sua attenzione sulla parte dell'input audio in cui è localizzato temporalmente il suono di interesse e sulle componenti in frequenza non influenzate dal rumore. Diversamente dal precedente DELayer, il

TF-DELayer utilizza un meccanismo di attenzione squeeze e excitation per ridurre drasticamente il costo computazionale dell'attenzione. L'analisi sperimentale ha dimostrato l'efficacia di DEGramNet, una CNN audio basata sul DEGram. DEGramNet è stata in grado di raggiungere risultati all'avanguardia sul dataset VGGSound (classificazione degli eventi sonori) e risultati comparabili con un approccio di ricerca dell'architettura di rete sul dataset VoxCeleb1 (identificazione degli speaker), dimostrando di essere un'architettura funzionale per diversi problemi di analisi audio. Inoltre, lo un'ulteriore analisi ha dimostrato che DEGram è più efficace dello spettrogramma, riuscendo a migliorare le prestazioni della stessa architettura CNN addestrata con entrambe le rappresentazioni e a ottenere una migliore accuratezza rispetto ai modelli più profondi. Infine, la riduzione del numero di features calcolate da DEGram consente alla stessa rete di ridurre il tempo di inferenza su architetture CPU, rendendo il modello in grado di funzionare in tempo reale su sistemi embedded. Per affrontare il problema della latenza percepita dall'utente, in questa tesi sono state proposte tre reti neurali multitask per eseguire contemporaneamente le seguenti attività: riconoscimento del genere, riconoscimento dell'emozione, stima dell'età e re-identificazione dell'interlocutore dalla sua voce. Ho dimostrato che è possibile non solo ridurre i requisiti computazionali rispetto ai corrispondenti singoli compiti, ma anche ottenere una migliore generalizzazione del modello stesso. L'esperimento ha permesso di identificare la migliore architettura da utilizzare in ambienti molto limitati e quando è richiesta una maggiore precisione. In particolare, ogni architettura multitask è caratterizzata da un diverso compromesso tra condivisione delle caratteristiche e parametri del modello. Tutte le architetture hanno in comune la CNN per l'estrazione delle features, ovvero ResNet18, e la rappresentazione audio, ovvero DEGram. Insieme ai modelli multitask, questa tesi propone anche una funzione di errore che consente di tenere conto di diversi dataset anche se non hanno le etichette per tutti i problemi da af-

frontare. Inoltre, è stato utilizzato l'algoritmo GradNorm per evitare uno squilibrio nelle prestazioni dei diversi problemi. Le architetture proposte sono state confrontate con algoritmi di ultima generazione e controparti single-task su benchmark standard per i problemi affrontati, ovvero VoxCeleb1 e VoxCeleb2 (Identificazione del parlante e stima dell'età), Mozilla Common Voice (Riconoscimento del genere) e IEMOCAP (Riconoscimento delle emozioni). I modelli addestrati hanno superato i modelli di ultima generazione in tre task su quattro, dimostrando la loro efficacia. Inoltre, l'approccio multitask ha dimostrato di migliorare le prestazioni del riconoscimento del genere, del riconoscimento delle emozioni e della stima dell'età rispetto ai modelli single task, convalidando l'ipotesi che il paradigma multitask non solo riduce i requisiti computazionali ma migliora anche le capacità di generalizzazione del modello.

Infine, questa tesi riduce ulteriormente la latenza di risposta del robot sociale identificando la migliore architettura Transformer per la comprensione del linguaggio naturale nel contesto della robotica sociale. In particolare, è stata valutata la compensazione tra l'accuratezza e il tempo di elaborazione di diverse architetture Transformer multitask. La valutazione tiene conto dei problemi legati all'indisponibilità di dati per l'addestramento del modello con un buon grado di generalizzazione e delle limitate risorse computazionali e di memoria dei sistemi embedded con GPU. Le architetture Transformer considerate sono caratterizzate da diverse approcci di ottimizzazione come la distillazione della conoscenza e le convoluzioni raggruppate. Da un lato, i risultati hanno dimostrato che utilizzando modelli Transformer è possibile ottenere buone prestazioni anche utilizzando un approccio basato su fine-tuning (indici di performance superiori al 90%). È stato anche dimostrato che i risultati sono fortemente influenzati dal problema affrontato (ad esempio, squilibrio delle frasi, dimensione del vocabolario e numero di entità). Per affrontare questo problema, nella tesi vengono identificate alcune strategie di progettazione che consentono di

scegliere meglio Transformer da utilizzare, in modo da evitare di spendere settimane ad addestrarli tutti. I tempi di elaborazione di questi modelli sono stati valutati su un sistema embedded comunemente utilizzato per le applicazioni robotiche, ovvero l'NVIDIA Jetson Xavier NX. L'analisi ha permesso di ridurre l'impatto del modello NLU sul tempo di risposta medio di circa il 20-25%. In generale, questa tesi esplora le architetture dei robot sociali in grado di essere empatici con gli esseri umani attraverso l'espressione di capacità di personalizzazione. Il prototipo proposto ha anche permesso di identificare le principali problematiche degli algoritmi di elaborazione dei dati sensoriali di ultima generazione per come sono percepiti dagli esseri umani. Questa tesi affronta i problemi identificati proponendo nuove rappresentazioni audio robuste al rumore ambientale e in grado di migliorare le prestazioni delle reti neurali leggere comunemente utilizzate sui robot sociali con un costo computazionale trascurabile. Inoltre, è stata affrontato il problema dell'ottimizzazione dei requisiti computazionali per gli algoritmi di analisi audio e NLU dei robot sociali proponendo due modelli multi-task in grado di ridurre la latenza di calcolo rispettivamente del 75% e del 33% rispetto ai modelli precedenti.

Abstract

Social robots are a particular category of robots able to perceive information about the environment to reason about the acquired information with the particular aim to interact with humans. Remarkable applications of social robots include museum guides, nurses, autism treatment, hotel assistants, and elderly care. Studies prove that the capability of social robots to personalize the conversation and perceive the interlocutor's emotions are the key behaviours that allow them to be considered emphatic.

To these purposes, social robots rely on multiple sensory modalities to acquire information about their interlocutor robustly and accurately. Their sensorial equipment is crucial considering that this kind of robot commonly works in challenging environments e.g., dynamic lighting conditions and loud environmental noise. In addition to these challenges, social robots must converse in real-time with humans to give them the feeling of natural interaction. Considering the application context, this result is only possible if the computation is performed on board of the social robot platform, which makes the task harder due to the implicit computational and power constraints.

This thesis tackles these requirements in the context of Deep Learning. In particular, a novel software architecture optimized for multimodal real-time interactions has been proposed as a general-purpose solution for social robots. The realization of a robotic prototype al-

lowed to identify the main issues perceived by humans about state-of-the-art algorithms related to human-robot interaction when deployed together in a real application. In light of this result, this thesis advances the state-of-the-art by proposing and validating novel auditory and natural language understanding algorithms optimized to be executed on robotic embedded systems while keeping high accuracy.

The proposed social robot architecture includes all the software modules that allow to meet the main requirements of a social robot: first, a dialogue manager able to personalize the human-robot interaction by exploiting the biometrics perceived by the sensors of the social robot; second, a multimodal sensor aggregation module able to exploits the information acquired by different types of sensors to increase the robustness to environmental noise; finally, parallel processing pipelines that, properly designed and implemented, ensure real-time performance. A social robot prototype based on the proposed architecture has been realized and deployed in the SICUREZZA exhibition for three days. 161 people who interacted with the robot evaluated their experience by answering 5 questions with a score between 1 and 5. The maximum score was achieved for more than 40% of the answers and the average rate was between 4 and 5. This result acquires more relevance considering that the people who attended the conference were technically skilled and, therefore, their feedback is reliable. The survey also allowed to investigate the feeling of humans about the performance of the state-of-art algorithms available on the proposed prototype. This analysis results in the need for audio algorithms more robust to environmental noise and more efficient human utterances processing pipelines.

Starting from this evidence, this thesis initially proposes two learnable audio representations to achieve the following goals: robustness to environmental noise and computational efficiency.

Firstly, a novel convolutional layer, namely Denoising-Enhancement Layer (DELayer), has been proposed. It is able to denoise and enhance

the input signal by combining an attention module (DELayer) with the SincNet layer; the final representation is thus able to attenuate the frequency components affected by environmental noise and amplify the ones relevant for the classification. The experimental results demonstrate that the end-to-end model based on the proposed DELayer, namely DENet, is definitely more effective than the existing methods in detecting and recognizing audio events of interest on the MIVIA Audio Events and on the MIVIA Road Events benchmarks. The novel CNN achieves state-of-the-art performance on both datasets and further analyses show its robustness to noise, its stability among different environmental conditions and its generalization capabilities.

The latency evaluation of the DENet model pointed out the need for a more optimized audio model to be run on CPU devices without batching the predictions and, therefore, increasing the speed of the model itself. For this reason, in this thesis the DELayer fundamentals have been extended to DEGram, a learnable spectrogram-based representation. Being a spectrogram-like representation, DEGram allows the use of shallower CNN with comparable accuracy. Consequently, it reduces the computational requirements of the final system while keeping the robustness of DENet. DEGram is the result of the combination of two novel layers: SincGram and a Time-Frequency version of the DELayer, namely TF-DELayer. The former, like SincNet, is able to learn the frequencies of interest for the audio analysis problem we are dealing with, extracting task-specific features through trainable band-pass filters. The latter can denoise the input signal from environmental background noise using a time-frequency attention mechanism, focusing on the part of the input signal in which the sound of interest is temporally located and on the frequency components not affected by noise. Differently from the previous DELayer, the TF-DELayer uses a squeeze and excitation attention mechanism to drastically reduce the attention overhead. The experimental analysis demonstrated the effectiveness of DEGramNet, an audio CNN based on the proposed

DEGram. DEGramNet was able to achieve state-of-the-art results on the VGGSound dataset (Sound Event Classification) and comparable results with a complex Network Architecture Search approach on the VoxCeleb1 dataset (Speaker Identification), proving to be a general-purpose architecture for audio analysis tasks. Moreover, the ablation study proved that the proposed DEGram representation was more effective than Spectrogram, being able to improve the performance of the same CNN architecture trained with both the representations and to achieve better accuracy with respect to deeper models. Finally, the reduction of the number of features computed by DEGram allows the same network to reduce the inference time on CPU architectures, making the model work in real-time on embedded systems.

To address the user-perceived latency issue, in this thesis three multitask neural network models have been proposed to perform the following tasks simultaneously: Gender Recognition, Emotion Recognition, Age Estimation, and Speaker Re-Identification. I proved that it is possible not only to reduce the computational requirements w.r.t. the single-task counterparts but also to obtain a better generalization of the model itself. The experiment allowed to identify the best architecture to use in very limited setups and when higher accuracy is required. In particular, each multitask architecture is characterized by a different trade-off between feature sharing and model parameters. All the architecture have in common the CNN backbone, i.e. ResNet18, and the audio representation, i.e. DEGram. Together with the multitask models, this thesis also proposes a training loss that allows to take into account different datasets even if they don't have the labels for all the tasks. Moreover, the GradNorm algorithm has been used to avoid unbalancing in the performance of the different tasks. The proposed architectures have been compared with state-of-the-art algorithms and single-task counterparts on standard benchmarks for the considered tasks, i.e. VoxCeleb1 and VoxCeleb2 (Speaker Re-Identification and Age Estimation), Mozilla Common Voice (Gender Recognition) and

IEMOCAP (Emotion Recognition). The trained models outperformed state-of-art models on three tasks over four proving their effectiveness. Moreover, the multitask approach proved to improve the performance of Gender Recognition, Emotion Recognition, and Age Estimation with respect to the single-task models, validating the hypothesis that the multitask paradigm not only reduces the computational requirements but also improves the generalization capabilities of the model.

Finally, this thesis further reduces the response latency of the social robot by identifying the best transformer architecture for Natural Language Understanding in the context of Social Robotics. In particular, the trade-off between the accuracy and processing time of different multitask transformer architectures has been evaluated. The evaluation takes into account the problems related to the unavailability of data for training the model with a good degree of generalization and the limited computational and memory resources of GPU-enabled embedded devices. The considered transformer architectures are characterized by different optimization approaches like knowledge distillation and grouped convolutions. On one hand, the results proved that using transformer models it is possible to achieve good performance even using a fine-tuning approach (performance indices over 90%). It has been also demonstrated that the results are highly influenced by the task (i.e., imbalance in the sentences, vocabulary size, and the number of entities). To address this issue, in the thesis are identified some design strategies allowing to better choice the transformer to use, so that it is possible to avoid wasting weeks training them all. The processing times of these transformers have been evaluated over an embedded system commonly used for robotic applications, namely the NVIDIA Jetson Xavier NX. The analysis allowed to reduce the impact of the NLU model on the average response time by around 20-25%.

Overall, this thesis explores Social Robot architectures able to be emphatic with humans through the expression of personalization capabilities. The proposed prototype also allowed to identify the main

gaps in state-of-art sensor processing algorithms as they are perceived by humans. This thesis also addresses the identified issues by proposing novel audio representations very robust to environmental noise and capable to gain the performance of shallow neural networks commonly running on Social Robots with a negligible computational cost. In addition, the optimization of the computational requirements for the audio analysis and NLU components of the social robots have been also addressed by proposing two multitask models able to reduce the computation latency by 75% and 33% with respect to earlier models, respectively.