Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 117-120 recibido 08-03-2017 revisado 11-05-2017 aceptado 22-05-2017

DeepVoice: Tecnologías de Aprendizaje Profundo aplicadas al Procesado de Voz y Audio

Deep Learning Technologies for Speech and Audio Processing

Marta R. Costa-jussà, José A. R. Fonollosa
TALP Research Center
Universitat Politècnica de Catalunya
Campus Nord, C/Jordi Girona, 08034 Barcelona
{marta.ruiz,jose.fonollosa}@upc.edu

Resumen: Este proyecto propone el desarrollo de nuevas arquitecturas para el procesado de la voz y el audio mediante métodos de aprendizaje profundo, explorando también nuevas aplicaciones y dando continuidad al trabajo inicial del equipo de investigadores solicitante y de toda la comunidad internacional. Las lineas de investigación incluyen: reconocimiento de voz, reconocimiento de eventos acústicos, síntesis de voz y traducción automática.
Palabras clave: Tecnologías del habla, aprendizaje profundo, eeconocimiento del habla, conversión de texto a voz, redes neuronales profundas

Abstract: This project proposes the development of new deep learning methods for speech and audio processing, exploring new applications and continuing the initial work of the research team and the international community. Research lines include: automatic speech recognition, acoustic event detection, speech synthesis and machine translation.
Keywords: Speech technology, deep learning, speech recognition, text to speech, deep neural networks

1 Participantes del proyecto

El grupo de investigación que participa en el proyecto es el grupo de Voz del Departamento de Teoría de Señal y Comunicaciones de la Universidad Politécnica de Cataluña. Los investigadores principales son los mismos autores de este artículo.

2 Entidad financiadora

El proyecto está financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional y el codigo del proyecto es TEC2015-69266-P. DeepVoice comenzó el 1 de enero de 2016 y tiene una duración de cuatro años.

3 Contexto y motivación

Las tecnologías de aprendizaje profundo hacen referencia a los métodos y sistemas de aprendizaje automático compuestos de varias capas de procesamiento o niveles de abstracción. Esta familia de algoritmos suele caracterizarse además por tener una estructura sencilla de describir y versátil. En concreto, este aprendizaje profundo suele utilizar alguna variante de las redes neuronales artificiales de múltiples capas o profundas para aprender un determinado modelo. En este modelado es tan importante la arquitectura de la red neuronal como el algoritmo de entrenamiento o aprendizaje de los parámetros de esta red.

En los últimos años, el modelado mediante redes neuronales ha resurgido con mucha fuerza gracias a ese énfasis en el aprendizaje y en el número de capas. Otros factores importantes han sido la disponibilidad de mayor capacidad de cálculo y de grandes bases de datos. La grandes bases de datos permiten entrenar mejor estructuras multicapa con gran número de parámetros y los recursos computacionales permiten realizar este proceso en tiempos razonables.

A pesar de que su uso no se ha generalizado hasta hace unos pocos años y de la dificultad de analizar el comportamiento de los algoritmos de aprendizaje profundo, su impacto ha sido ya espectacular en mucho ámbi-

ISSN 1135-5948

tos como el procesado de imagen, voz y texto tanto a nivel de investigación como comercial. En reconocimiento de voz, por ejemplo, se ha pasado de un avance anual muy lento basado en sistemas de gran complejidad a estructuras sencillas de aprendizaje profundo que suponen toda una revolución en cuanto a arquitectura y salto en prestaciones.

Este proyecto propone el desarrollo de nuevas arquitecturas para el procesado de la voz y el audio mediante métodos de aprendizaje profundo, explorando también nuevas aplicaciones.

El proyecto incluye un paquete de trabajo general dedicado al aprendizaje profundo y otros cuatro paquetes de trabajo dedicados al reconocimiento del habla y del locutor, detección de eventos acústicos, síntesis de voz y traducción de voz. En el primer paquete de trabajo se exploran nuevas arquitecturas y algoritmos de aprendizaje, teniendo en cuenta el coste computacional y la escalabilidad a grandes bases de datos, mientras que los siguientes exploran su aplicación en procesado de la voz y del audio. En la siguiente sección mencionamos con algo más de detalle qué aportaciones se harán en cada una de las tareas.

En estas tareas o en la difusión de los resultados está previsto continuar colaborando con otros grupos de investigación a nivel nacional e internacional y con las empresas interesadas en la temática del proyecto y sus resultados. En concreto, se incluye en el plan de trabajo la colaboración con el hospital Sant Joan de Déu de Barcelona en la detección y mejora de las condiciones acústicas de las unidades de cuidados intensivos de neonatos. También se pone énfasis en la evaluación de los resultados. Se comenta esta colaboración en la sección 5 de este artículo.

4 Proyecto DeepVoice

El proyecto integra diferentes áreas de las tecnologías del habla y prentende contribuir en cada una de ellas incorporando modelos de aprendizaje profundo. A continuación describimos brevemente los objetivos de cada uno de los paquetes de trabajo del proyecto que además del paquete de arquitecturas de aprendizaje profundo incluye las areas de: reconocimiento de voz, reconocimiento de eventos acústicos, síntesis de voz y traducción automática.

4.1 Arquitecturas de aprendizaje profundo

Las arquitecturas profundas construidas a partir de redes neuronales artificiales tienen una larga historia, pero su reciente renacimiento está relacionado con la disponibilidad de algoritmos de entrenamiento eficaces, bases de datos grandes y hardware de computación potente (Hinton, Osindero, y Teh, 2006; Bengio, 2009).

El proyecto dedicará recursos a investigar nuevas arquitecturas de aprendizaje profundo que puedan ser útiles en aplicaciones de voz. Se pretende desarrollar medidas de optimización nuevas para entrenar redes recurrentes con datos no segmentados. Asimismo, desarrollar nuevos algoritmos de entrenamiento o modificar los ya existentes para que sean paralelizables.

4.2 Reconocimiento de voz

El impacto del aprendizaje profundo en reconocimiento de voz ha sido revolucionario y abarcan las tres líneas de investigavión que vamos a seguir en este proyecto.

En primer lugar, en robustez del sistema de reconocimiento, algunos trabajos recientes proponen usar redes neuronales profundas (Xia y Bao, 2014) para reducir el ruido de la señal, por poner un ejemplo. En esta dirección, se contribuirá mediante el desarrollo de técnicas basadas en aprendizaje profundo que permitan añadir ruido al sistema sin que la calidad se vea afectada.

En segundo lugar, se pretende desarrollar arquitecturas *end-to-end* de reconocimiento de voz, viendo la viabilidad de las mismas en ejemplos anteriores (Hannun et al., 2014). Para ello, se debe hacer un estudio exhaustivo de las características perceptuales en modelado acústico y su modelización con modelos neuronales profundos. Asimismo, se pretende usar redes neuronales recurrentes y entrenamientos conjuntos para los modelos acústico y de lenguaje.

Finalmente, en reconocimiento de locutor trabajos anteriores como (Richardson, Reynolds, y Dehak, 2015) usan las redes neuronales para extracción automática de características. En este proyecto se pretende ir más allá y usar la entrada de señal sin modificar para mejorar el rendimiento de los algoritmos de aprendizaje profundo.

4.3 Reconocimiento de eventos acústicos

El contexto de esta tarea se encuentra en la unidad de curas intensivas de neonatos (NICU). En este contexto, hay muchos ruidos que se tienen que filtrar para estudiar los patrones relevantes. Se pretende grabar y etiquetar datos recogidos de micrófonos instalados en las incubadoras de las NICU. La base de datos incluirá información sobre las variables fisiológicas relevantes y los patrones de sueño.

4.4 Síntesis de voz

El aprendizaje profundo se ha integrado en síntesis de voz principalmente aplicado a la modelización paramétrica (Ling et al., 2015)

La tarea de síntesis de voz es básicamente una tarea de regresión. Con tal de producir voz natural y continua se pueden utilizar técnicas de generación paramétrica. En esta area, proponemos investigar representaciones de la voz que permitan usar redes neuronales. También pretendemos proponer y evaluar técnicas de aprendizaje profundo para reducir el ruido de la voz generada e incluir expresividad en la voz final.

4.5 Traducción automática

En este caso, el aprendizaje profundo se ha usado para mejorar los sistemas estadísticos ya existentes y también ha permitido desarrollar un nuevo paradigma de traducción usando un modelado de secuencia a secuencia. Como en las otras areas, la lista de trabajos es muy extensa (Costa-jussà et al., 2017).

La traducción automática se puede aplicar a la voz o al texto. El objetivo al final de este proyecto es construir un sistema de traducción de voz a texto, ya sea concatenando técnicas de reconocimiento de voz y traducción de texto o planteando un sistema directo de voz a texto traducido. En el primer caso, se integrarán las mejoras del paquete de reconocimiento de voz y las mejoras que aporta un paradigma de traducción automática basado en redes neuronales. En el segundo caso, se diseñará una nueva arquitectura neuronal para afrontar el reto.

5 Impacto del proyecto

Las tecnologías de voz pueden facilitar el acceso a la información (comunicación hombre-máquina) y la comunicación humana. Los dispositivos electrónicos se están convirtiendo en imprescindibles. El uso de la voz en estos dispositivos es cada vez más esencial y también puede abrir una nueva gama de posibilidades. Estas tecnologías también pueden aplicarse a múltiples campos específicos, como mejorar la comunicación y la comprensión de los seres humanos, ayudar a las personas discapacitadas y ancianas, mejorar los servicios ofrecidos en los medios de comunicación, etc. El empleo de dispositivos de voz con voces inadecuadas (género, edad, acento, dialecto, tono) o sistemas de reconocimiento de voz que no funcionan en condiciones ruidosas pueden desalentar a los usuarios. El desarrollo que estamos proponiendo de la tecnología de voz será la clave para aplicaciones robustas de alta calidad. Asimismo, la traducción es un aspecto importante para reducir las barreras internacionales y lograr el pleno entendimiento entre las personas, preservando al mismo tiempo las sociedades multilingües. Esperamos realizar traducciones de voz en tiempo real y de alta calidad con concatenación e integración de reconocimiento profundo de voz y tecnologías de traducción automática. Esto representaría un progreso claro en los negocios y las relaciones políticas, así como en las áreas de ocio y educación.

Nuestra propuesta de investigación sobre detección de eventos acústicos también incluye su aplicación específica en unidades de cuidados intensivos neonatales (NICU). En este caso, se diferenciarán los factores de ruido microambiental y los signos fisiológicos y así los clínicos podrán proponer mejores protocolos NICU.

6 Página web

En la página web del proyecto
http://www.tsc.upc.edu/deepvoice/
se puede consultar el equipo de investigación. En la misma página también se harán públicos los principales resultados alcanzados con el progreso de DeepVoice.

Bibliografía

Bengio, Y. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, Enero.

Costa-jussà, M. R., A. Allauzen, L. Barrault, K. Cho, y H. Schwenk. 2017. Introduction to the Special Issue on Deep Learning Approaches for Machine Translation. *Accepted for publication in Computer Speech*

and Language, Special Issue in Deep learning for Machine Translation.

Hannun, A. Y., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, y A. Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

Hinton, G. E., S. Osindero, y Y. Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, Julio.

Ling, Z., S. Kang, H. Zen, A. W. Senior, M. Schuster, X. Qian, H. M. Meng, y L. Deng. 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag.*, 32(3):35–52.

Richardson, F., D. A. Reynolds, y N. Dehak. 2015. A unified deep neural network for speaker and language recognition. *CoRR*, abs/1504.00923.

Xia, B. y C. Bao. 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, 60:13–29.