Reconocimiento Automático del Habla
2023-2024

Deep Learning en RAH

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

MIARFID
Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

MIARFID-RAH mcastro@dsic.upv.es

¿Por qué reconocimiento del habla *ahora*?

- El *reconocimiento del habla* está invadiendo nuestras vidas. En nuestro móvil (Siri), en los videojuegos (Kinect), en los relojes (Apple Watch), en nuestros hogares (Alexa). ¿Por qué ***ahora***?

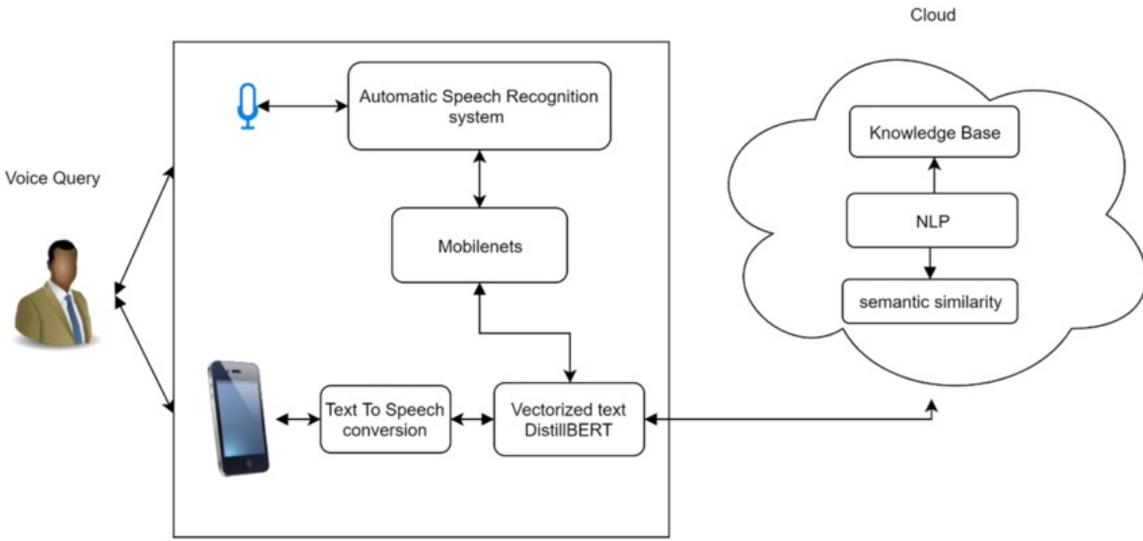
- Lo que ha ocurrido es que el *Deep Learning* ha hecho posible que los sistemas de reconocimiento sean lo suficiente buenos para usarlos en entornos abiertos.

Cómo era el mundo en 2005…

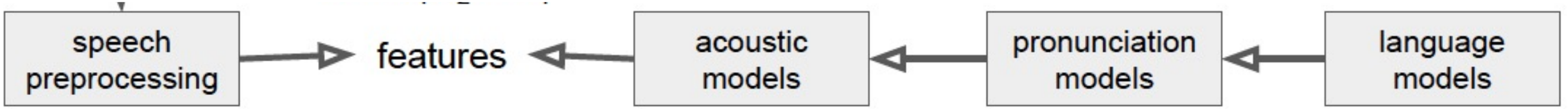
- Google search engine muy *popular*... en nuestros **ordenadores**
- Los móviles (aún llamados móviles, no “smartphones”) también eran muy *populares* pero solo para *hablar por* **teléfono** (nada de acceso a internet)
- El RAH ya existía... pero como aplicaciones al gran público básicamente era **para no hablar contigo**: “Diga 1 si quiere darse de baja, diga 2 si quiere darse alta...”
- No nos gustaba el RAH y pensábamos que no era útil.
- No pensaban así las grandes corporaciones que veían ya un futuro muy lucrativo incorporando RAH en los móviles y en otras aplicaciones como el subtitulado automático de vídeos (y su traducción a otras lenguas).

y en 2007… La revolución

- En 2007 se lanzaron al mercado los primeros *smartphones*
  - https://en.wikipedia.org/wiki/IPhone_(1st_generation)
  - https://en.wikipedia.org/wiki/Android_(operating_system)
- y nos empezó a gustar el RAH
- y empezamos a utilizarlo
- y empezamos a generar grandes cantidades de datos



¿Cómo es un sistema de RAH?

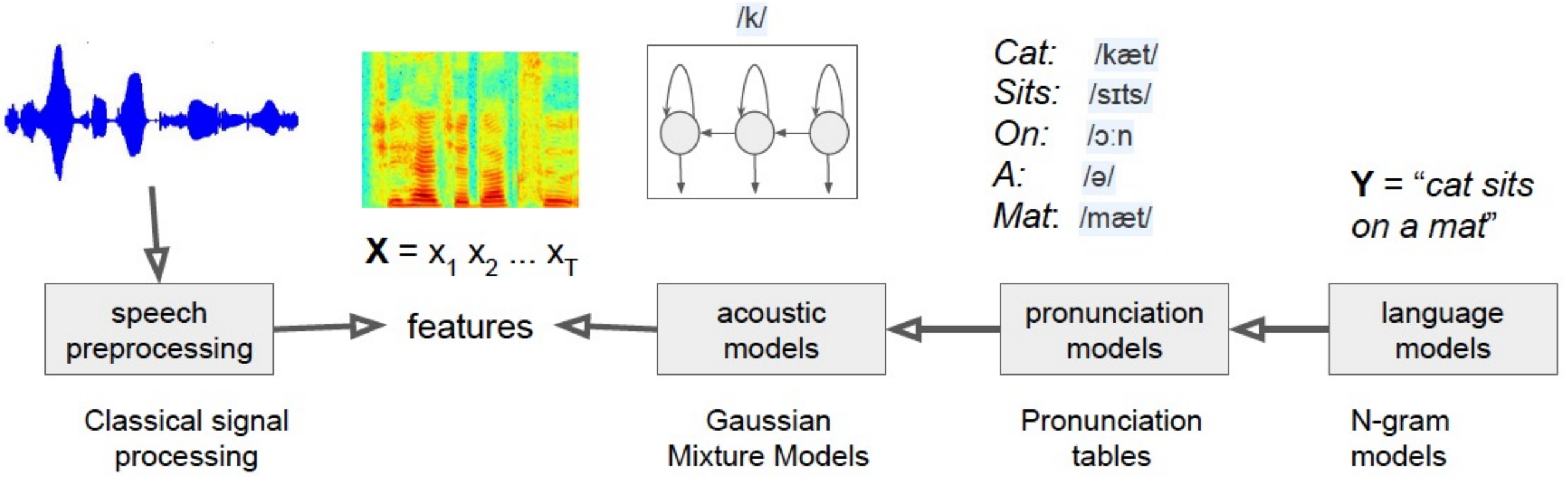


Modelo de lenguaje: *¿cómo decimos algo?*

Modelo de pronunciación o Léxico: *¿cómo lo pronunciamos?*

Modelo acústico: *correspondencia entre señal acústica y secuencia de palabras*

El enfoque clásico



- Todos estos modelos se **entrenan de forma independiente con datos separados** y
- se **combinan** para inferir

→la secuencia más probable de palabras correspondientes a la entrada acústica.

Vamos a revisar este enfoque clásico para entender la evolución a los sistemas actuales de RAH

¿Qué es “Deep Learning”?

- A sub-field within machine learning based on algorithms for learning multiple levels of representation in order to model complex relationships among data.
- Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture.

Deep Learning: técnicas de *Machine Learning*, generalmente basadas en RNAs que pueden aprender múltiples niveles de representación para modelar relaciones complejas.

Deep Learning

Modern Reincarnation of Artificial Neural Networks

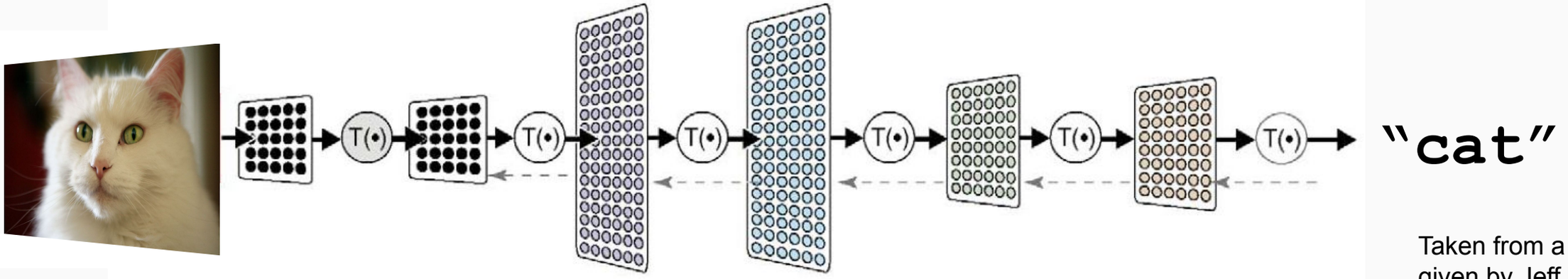
Collection of simple trainable mathematical units, organized in layers, that work together to solve complicated tasks

What's New

new network architectures,
new training math, ***scale***

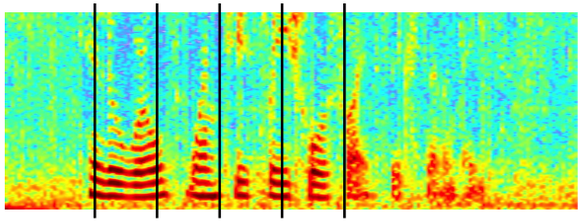
Key Benefit

Learns features from raw, heterogeneous, noisy data
No explicit feature engineering required



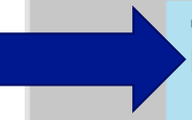
Taken from a presentation given by Jeff Dean, 2017

Functions a Deep Neural Network Can Learn

Input		output
Pixels:		“lion”
Audio:		“How cold is it outside?”
“Hello, how are you?”		“Bonjour, comment allez-vous?”
Pixels:		“A blue and yellow train travelling down the tracks”

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013



Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. →

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous. →

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child? →

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts. →

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people. →

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss. →

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket. →

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible. →

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases. →

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical. →

The New York Times

Monday, June 25, 2012 Last Update: 11:50 PM ET

Arizona Ruling Only a Narrow Opening for Other States

By JULIA PRESTON 9:06 PM ET

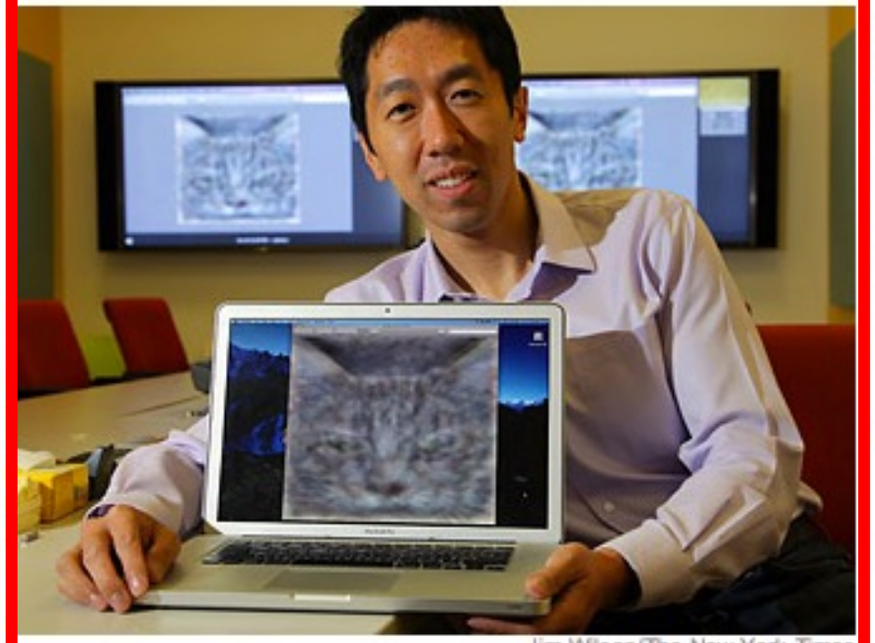
The Supreme Court's mixed decision on Arizona's immigration law does not seem likely to unleash a new wave of legislation by other states to crack down on illegal immigration.

- Blocking Parts of Arizona Law, Justices Uphold Centerpiece
- On Campaign Trail, Obama and Romney React 10:05 PM ET
- Interactive: Supreme Court Decision on Immigration Law

Justices Bar Mandatory Life Terms for Juveniles

By ADAM LIPTAK and ETHAN BRONNER 8:44 PM ET

The justices ruled that such sentencing for those under 18



Jim Wilson/The New York Times

Despite Itself, a Simulated Brain Seeks Cats

By JOHN MARKOFF 12 minutes ago

A Google research team, led by Andrew Y. Ng, above, and Jeff Dean, created a neural network of 16,000 processors that reflected human obsession with Internet felines.

Basketball Recruiters Aim for Middle Schoolers

By ADAM HIMMELSBACH and PETE THAMEL 10:03 PM ET

OPINION »

OPINIONATOR | ANXIETY

Haunted Heart

The surgeon calmly fiddled with the beating engine of my existence.

- Krugman: Abdication
- Edsall: Working-Class Chaos
- Editorial: Spending Cuts
- The Stone: Evolution and Our Inner Conflict
- Taking Note: Citizens United

MARKETS »

At 11:48 PM ET

JAPAN		CHINA
Nikkei	HangSeng	Shanghai
8,660.36	18,901.11	2,214.11
–74.26	+3.66	–10.00
–0.85%	+0.02%	–0.45%

Data delayed at least 15 minutes

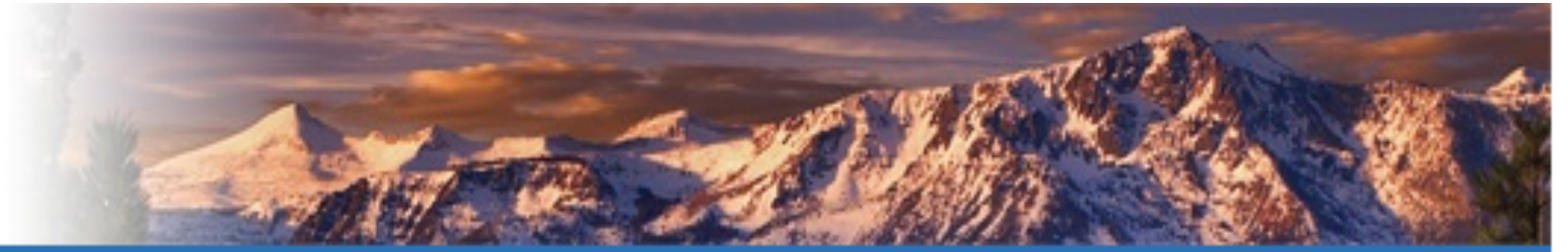
GET QUOTES My Portfolios »

Stock, ETFs, Funds Go

Rick Rashid in **Tianjin**, October, 25, 2012 https://www.youtube.com/watch?v=Nu-nlQqFCKg

La “industria” y la “academia”

Neural Information Processing Systems Foundation



NIPS : Conferences : 2009 : Program

- NIPS Home
- Overview
- Conference Videos
- Workshop Videos
- Program Highlights
- Tutorials
- Conference Sessions
- Workshops
- Publication Models
- Demonstrations
- Mini Symposia

Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

Deep Learning for Speech Recognition and Related Applications

7:30am - 6:30pm Saturday, December 12, 2009

Location: Hilton: Cheakamus

Abstract: Over the past 25 years or so, speech recognition technology has been dominated by a “shallow” architecture --- hidden Markov models (HMMs). Significant technological success has been achieved using complex and carefully engineered variants of HMMs. The next generation of the technology requires solutions to remaining technical challenges under diversified deployment environments. These challenges, not adequately addressed in the past, arise from the many types of variability present in the speech

Publicaciones conjuntas

IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups

INTERSPEECH 2010

Binary Coding of Speech Spectrograms Using a Deep Auto-encoder

L. Deng[1], M. Seltzer[1], D. Yu[1], A. Acero[1], A. Mohamed[2], and G. Hinton[2]

[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, US
[2] University of Toronto, Toronto, Ontario, Canada
{deng|mseltzer|dongyu|alexac}@microsoft.com; {asamir|hinton}@cs.toronto.edu

Abstract

This paper reports our recent exploration of the layer-by-layer learning strategy for training a multi-layer generative model of patches of speech spectrograms. The top layer of the

The work reported in this paper was inspired by the successful use of deep auto-encoders for dimensionality reduction [8][9] and the extension of this work to the discovery of efficient binary codes in information retrieval [12]. It is also motivated by the potential benefits of using

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Yann LeCun: Facebook AI Research, New York University
Yoshua Bengio: Université de Montréal
Geoffrey Hinton: Google, University of Toronto

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.

Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to beating records in image recognition[1–4] and speech recognition[5–7], it has beaten other machine-learning techniques at predicting the activity of potential drug molecules[8], analysing particle accelerator data[9,10], reconstructing brain circuits[11], and predicting the effects of mutations in non-coding DNA on gene expression and disease[12,13]. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding[14], particularly topic classification, sentiment analysis, question answering[15] and language translation[16,17].

We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress.

Supervised learning

The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large data set of images of houses, cars, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input–output function of the machine. In a typical deep-learning system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine.

To properly adjust the weight vector, the learning algorithm computes a gradient vector that, for each weight, indicates by what amount the error would increase or decrease if the weight were increased by a tiny amount. The weight vector is then adjusted in the opposite direction to the gradient vector.

The objective function, averaged over all the training examples, can

[1]Facebook AI Research, 770 Broadway, New York, New York 10003, USA. [2]New York University, 715 Broadway, New York, New York 10003, USA. [3]Department of Computer Science and Operations

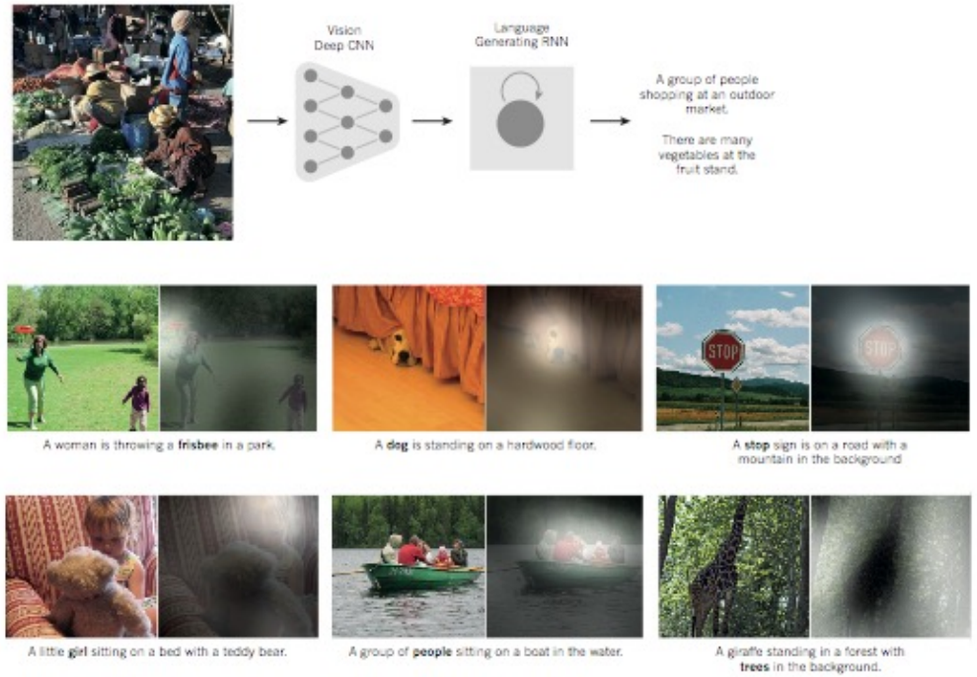


Figure 3 | From image to text. Captions generated by a recurrent neural network (RNN) taking, as extra input, the representation extracted by a deep convolution neural network (CNN) from a test image, with the RNN trained to 'translate' high-level representations of images into captions (top). Reproduced with permission from ref. 102. When the RNN is given the ability to focus its attention on a different location in the input image (middle and bottom; the lighter patches were given more attention) as it generates each word (bold), we found[86] that it exploits this to achieve better 'translation' of images into captions.

self-driving cars[60,61]. Companies such as Mobileye and NVIDIA are using such ConvNet-based methods in their upcoming vision systems for cars. Other applications gaining importance involve natural language understanding[14] and speech recognition[7].

Despite these successes, ConvNets were largely forsaken by the mainstream computer-vision and machine-learning communities until the ImageNet competition in 2012. When deep convolutional networks were applied to a data set of about a million images from the web that contained 1,000 different classes, they achieved spectacular results, almost halving the error rates of the best competing approaches[1]. This success came from the efficient use of GPUs, ReLUs, a new regularization technique called dropout[62], and techniques to generate more training examples by deforming the existing ones. This success has brought about a revolution in computer vision; ConvNets are now the dominant approach for almost all recognition and detection tasks[4,58,59,63–65] and approach human performance on some tasks. A recent stunning demonstration combines ConvNets and recurrent net modules for the generation of image captions (Fig. 3).

Recent ConvNet architectures have 10 to 20 layers of ReLUs, hundreds of millions of weights, and billions of connections between units. Whereas training such large networks could have taken weeks only two years ago, progress in hardware, software and algorithm parallelization have reduced training times to a few hours.

The performance of ConvNet-based vision systems has caused most major technology companies, including Google, Facebook, Microsoft, IBM, Yahoo!, Twitter and Adobe, as well as a quickly growing number of start-ups to initiate research and development projects and to deploy ConvNet-based image understanding products and services.

ConvNets are easily amenable to efficient hardware implementations in chips or field-programmable gate arrays[66,67]. A number of companies such as NVIDIA, Mobileye, Intel, Qualcomm and Samsung are developing ConvNet chips to enable real-time vision applications in smartphones, cameras, robots and self-driving cars.

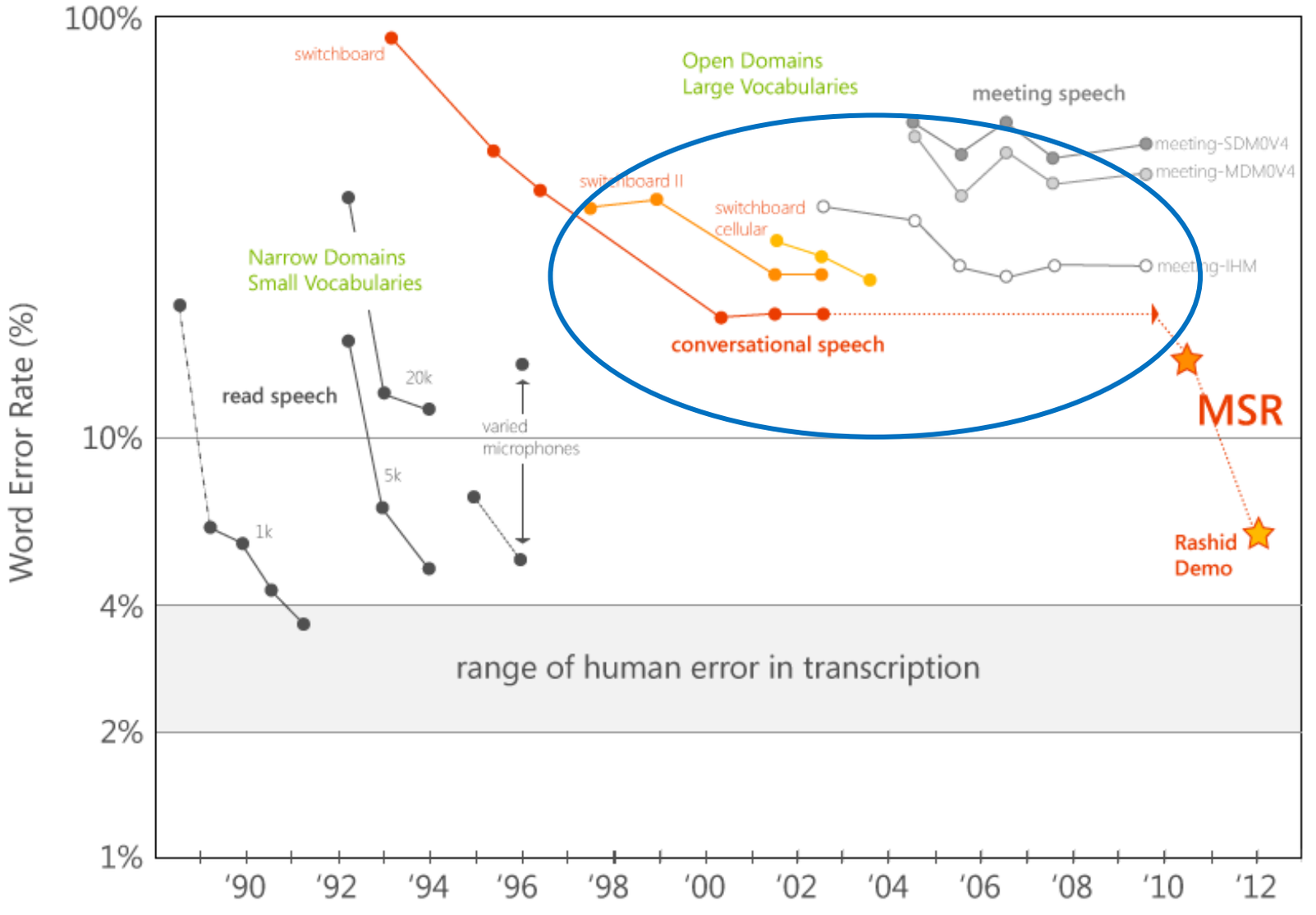
Distributed representations and language processing

Deep-learning theory shows that deep nets have two different exponential advantages over classic learning algorithms that do not use distributed representations[21]. Both of these advantages arise from the power of composition and depend on the underlying data-generating distribution having an appropriate componential structure[40]. First, learning distributed representations enable generalization to new combinations of the values of learned features beyond those seen during training (for example, 2^n combinations are possible with n binary features)[68,69]. Second, composing layers of representation in a deep net brings the potential for another exponential advantage[70] (exponential in the depth).

The hidden layers of a multilayer neural network learn to represent the network's inputs in a way that makes it easy to predict the target outputs. This is nicely demonstrated by training a multilayer neural network to predict the next word in a sequence from a local

De izquierda a derecha Russ Salakhutdinov, Richard S. Sutton, Geoffrey Hinton, Yoshua Bengio y Steve Jurvetson en 2016

Evaluaciones NIST

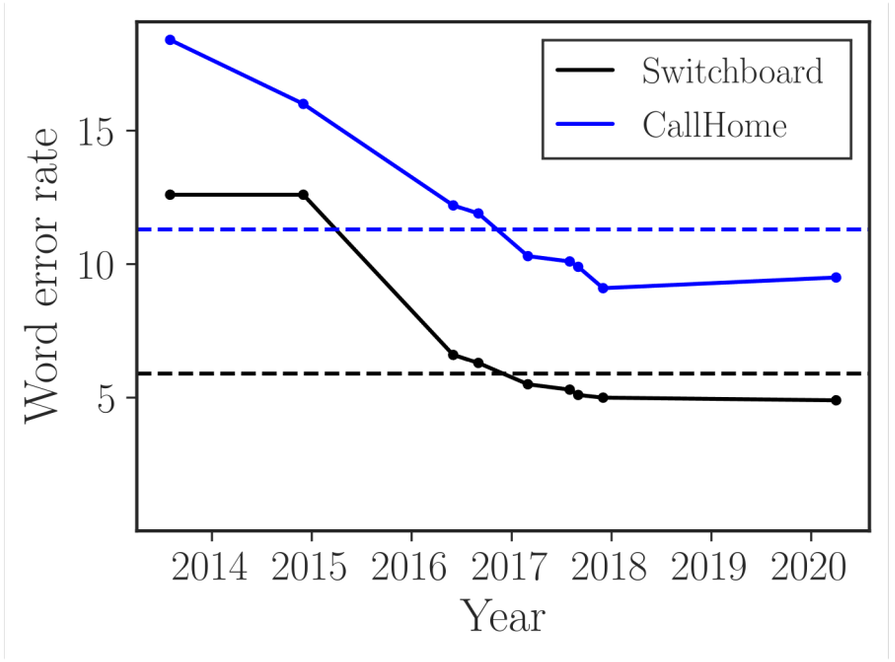


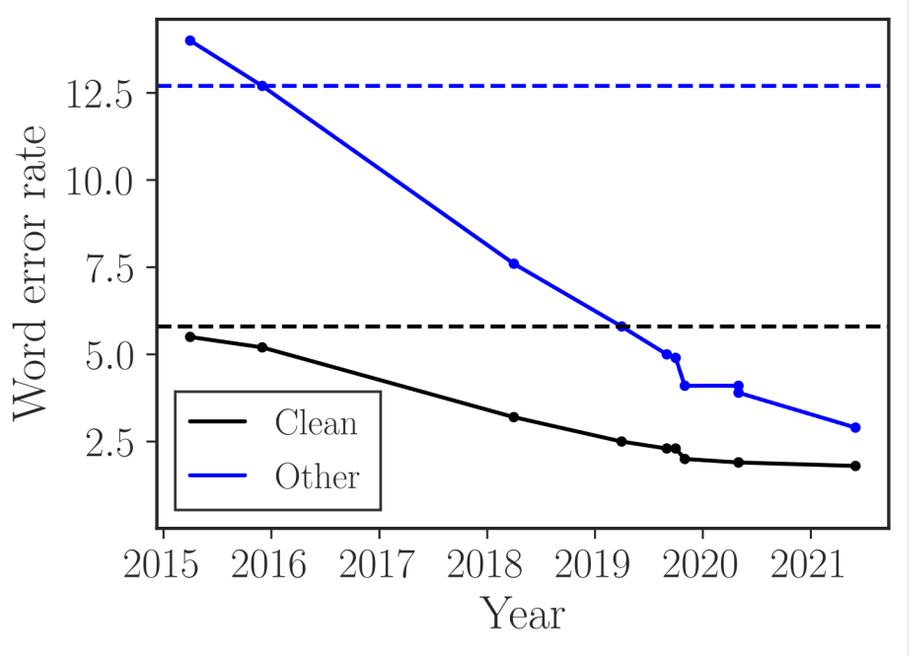
Después de que la comunidad científica no mejorase los resultados durante más de 10 años….

…Microsoft Research redujo el error de **~23%** a **<15%** (y menos del 7% en la demo de Rick)

Li Deng (MSRR), Dong Yu (MSRR),
Geoffrey Hinton (Univ Toronto),
Frank Seide (MSRA)

Evaluaciones NIST



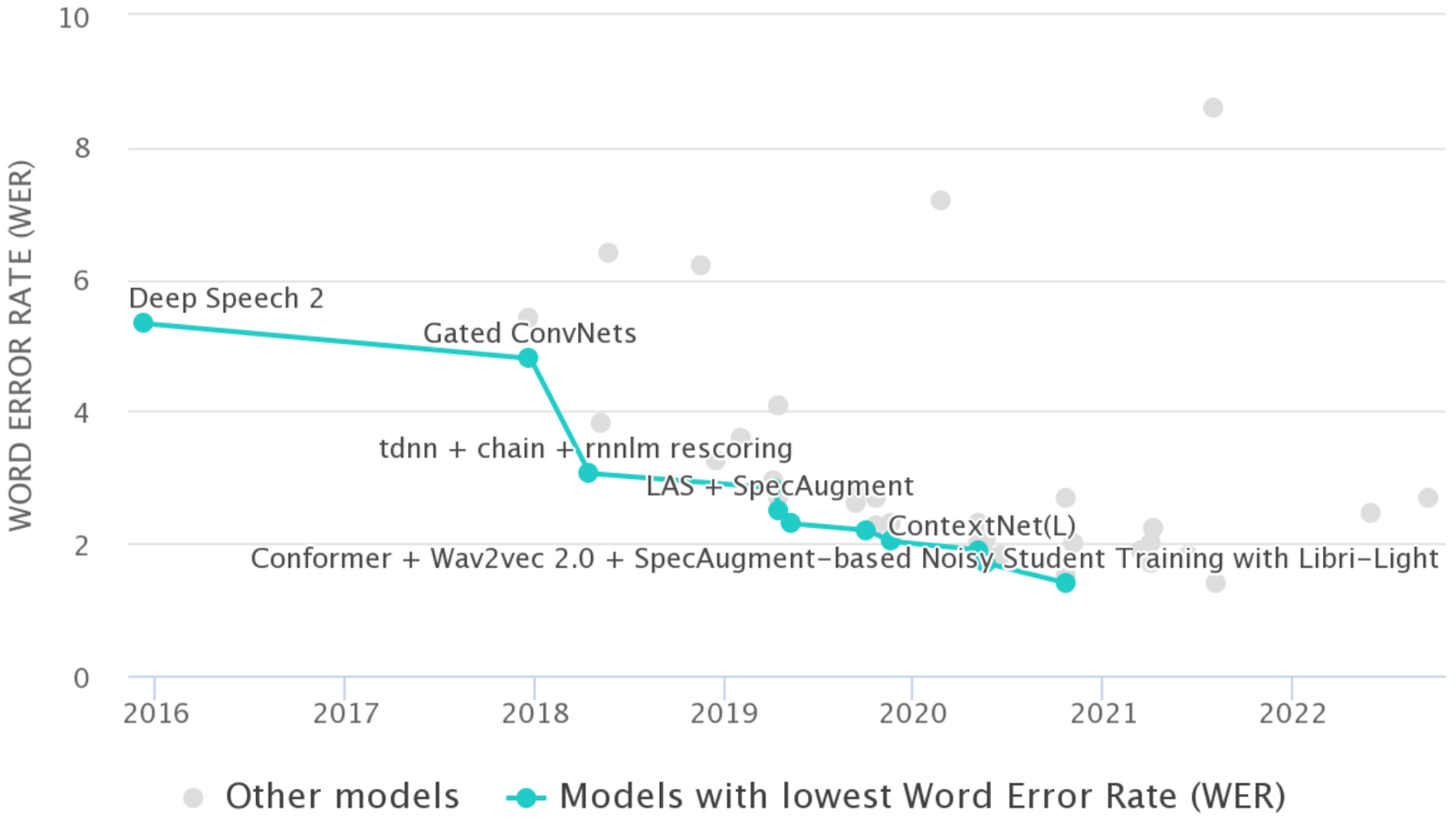


WER en Switchboard Hub5'00 (izquierda) y LibriSpeech (derecha). Las líneas discontinuas indican el nivel humano. arXiv:2108.00084

Evaluaciones en RAH

Corpus LibriSpeech

- Aproximadamente 1,000 horas de grabaciones de audio
- Habla “leída” de audiolibros
- Código abierto en 2015
- Carrera altamente competitiva liderada por Google, Facebook y ASAPP



https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

Evaluaciones en RAH

	Rank	Model	Word Error Rate (WER)	Extra Training Data	Paper	Code	Result	Year	Tags
Google	1	**Conformer + Wav2vec 2.0 + SpecAugment-based Noisy Student Training with Libri-Light**	1.4	✓	Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition			2020	Conformer
Google	2	**w2v-BERT XXL**	1.4	✓	W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training			2021	
facebook	3	**Conv + Transformer + wav2vec2.0 + pseudo labeling**	1.5	✓	Self-training and Pre-training are Complementary for Speech Recognition			2020	Transformer
Google	4	**ContextNet + SpecAugment-based Noisy Student Training with Libri-Light** ed3book_jan122022.pdf		✓	Improved Noisy Student Training for Automatic Speech Recognition			2020	LSTM
Google	5	**SpeechStew** (1B)	1.7	×	SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network			2021	Conformer
ASAPP	6	**Multistream CNN with Self-Attentive SRU**	1.75	×	ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition			2020	CNN

https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

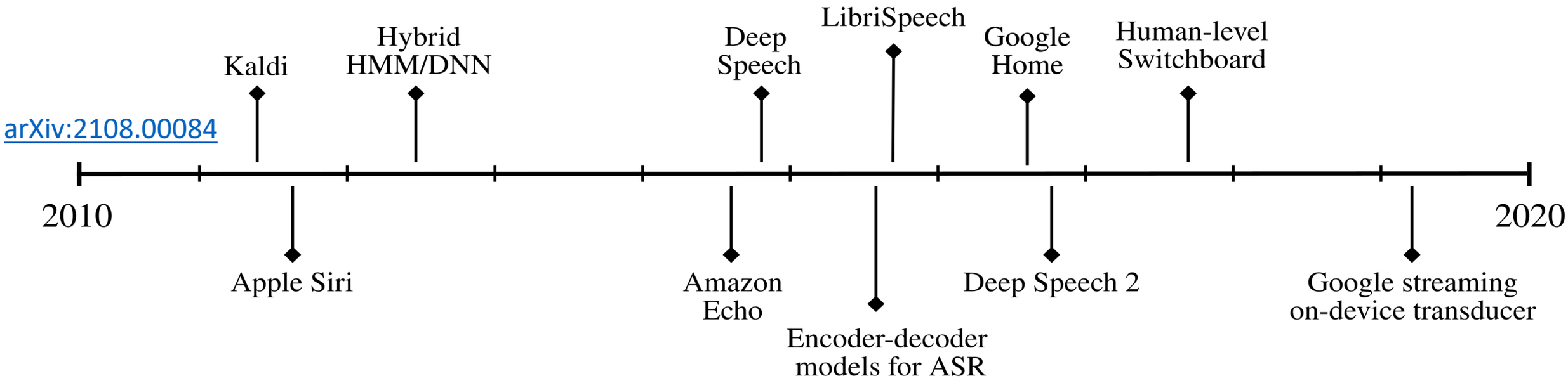
Evaluaciones en RAH

English Tasks	WER%
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

Figure 26.1 Rough Word Error Rates (WER = % of words misrecognized) reported around 2020 for ASR on various American English recognition tasks, and character error rates (CER) for two Chinese recognition tasks.

(De: Speech and Language Processing. Daniel Jurafsky & James H. Martin. 2021)

Cronología RAH 2010-2020



La década ha visto:

- el lanzamiento de dispositivos basados en voz y asistentes de voz (Siri, Alexa, Google Home...),
- software de reconocimiento de voz de código abierto y ampliamente utilizado como Kaldi o Whisper, y
- corpus de referencia muy grandes como LibriSpeech.

Claves del éxito del Deep learning en RAH:

1) la creación de corpora de datos transcritos masivos,
2) la mejora del hadware (GPUs),
3) la mejora en los algoritmos de aprendizaje y arquitecturas.

4 Generaciones en RAH

- **Generación 1 (1950s-1960s)**

  Sistemas basados en acústica y fonética
- **Generación 2 (1960s-1970s)**

  Sistemas basados en plantillas
- **Generación 3 (1970s-2000s)**

  Sistemas basados en modelado estadístico
- **Generación 4 (2000s-present)**

  Sistemas basados en Deep Learning

En la industria, el estado del arte cambió en los últimos años de

HMM-GMM + N-grams

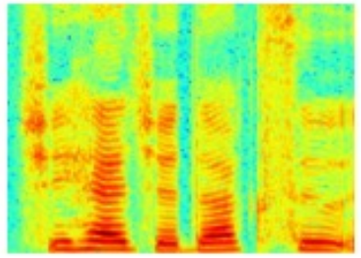
a

Deep (forward) neural network + Recurrent neural network

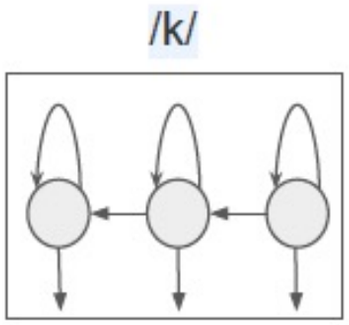
CNN (feature extraction from raw signal) + LSTM, con más fuentes de información

End-to-end approaches (sequence-to-sequence classifiers)

El enfoque clásico



$\mathbf{X} = x_1\ x_2\ \ldots\ x_T$



Cat: /kæt/
Sits: /sɪts/
On: /ɔːn
A: /ə/
Mat: /mæt/

Y = "*cat sits on a mat*"

speech preprocessing → features ← acoustic models ← pronunciation models ← language models

Classical signal processing

Gaussian Mixture Models

Pronunciation tables

N-gram models

- Todos estos modelos se **entrenan de forma independiente con datos separados** y
- se **combinan** para inferir

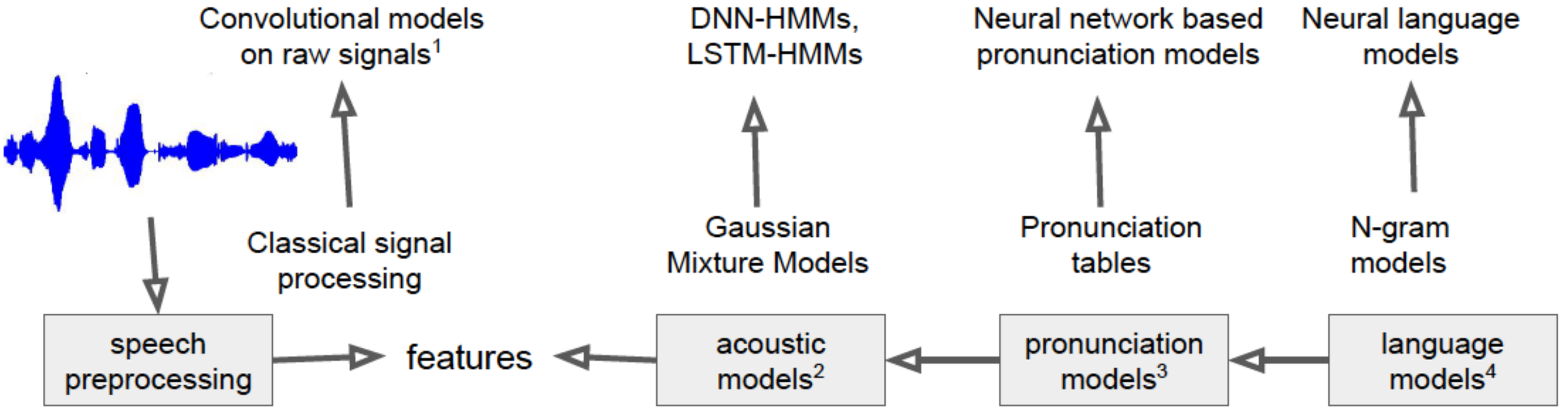
→la secuencia más probable de palabras correspondientes a la entrada acústica.

La invasión de las RNAs

Los investigadores se iban dando cuenta que cada uno de los componentes de un sistema de RAH podía funcionar de manera más efectiva con RNAs:

- ML de N-gramas → ML conexionista
- Modelos de pronunciación → podemos descubrir cómo pronunciar una nueva secuencia de caracteres que nunca antes habíamos visto usando una RNA
- Modelos acústicos → Deep ANN (LSTM)
- Preproceso → se podía reemplazar por redes neuronales convolucionales en señales de voz sin procesar.

Componentes basados en RNAs



- Ya hay **RNAs en cada componente**, pero aún se **entrenan con datos separados** y
- se **combinan** para inferir

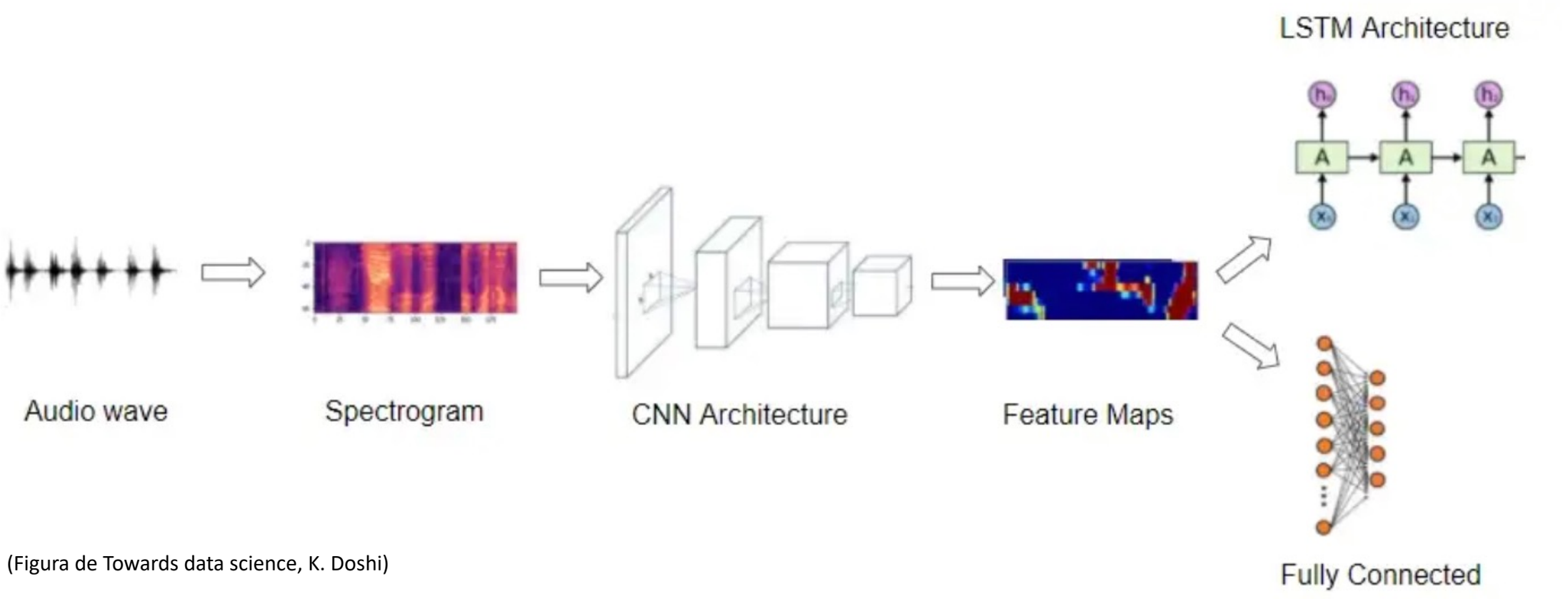
→la secuencia más probable de palabras correspondientes a la entrada acústica.

Modelado Deep

- Deep Preprocessing
  - ➢ deep feedforward neural networks (CNN)
- Deep Acoustic Modeling
  - ➢ deep feedforward neural networks (DNN, CNN, $\cdots$)
  - ➢ deep recurrent neural networks (LSTM, GRU, $\cdots$)
  - ➢ end-to-end models
- Deep Language Modeling

Deep Preprocessing

- No necesidad de utilizar las técnicas tradicionales de procesamiento de señal.
- El enfoque con Deep Learning es utilizar los datos de audio en su forma original, PERO convirtiéndolos en “imágenes” (espectogramas) y luego usar una arquitectura CNN estándar para procesar esas imágenes.
- Recordemos que un espectrograma es una representación compacta de la señal de audio, algo así como una “huella digital” de la señal.



(Figura de Towards data science, K. Doshi)

Deep Preprocessing

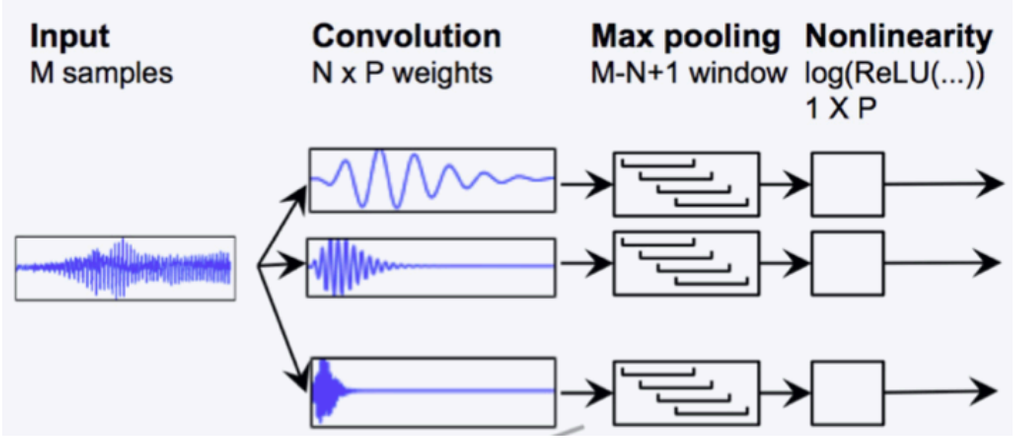
Método: A partir de los datos de audio *raw*

1. Calcular espectograma
2. Opcionalmente, utilizar técnicas simples de procesamiento de señal para alimentar también al sistema. (También se puede realizar algún tipo de limpieza de la señal antes de calcular el espectograma)
3. Una vez tenemos el espectograma (la **imagen**), podemos usar arquitecturas CNN estándar para procesarlos y extraer mapas de características que son una representación codificada de la imagen del espectrograma.

Deep Preprocessing

Using deep neural net classifiers, filters directly from speech signal

Sainath et al ASRU 2013



DNN-based ASR

Q: what are the learned weights in the convolution input layer?
A: impulse responses of filters consistent with critical bands of hearing

also Palatz et al 2013, Tueske et al 2014, Golik et al 2015, Gharemani et al 2016, Luo and Mesgarani 2918, …

From Hynek Hermansky (ASRU 2019)

Modelos end-to-end

¿Qué es?

Dado

- un audio *X* (la señal o una secuencia de frames, de talla *T*) y
- la correspondiente salida *Y* (secuencia de palabras correspondiente, de talla *N*),

el sistema de RAH es un modelo probabilístico

$$p(Y|X)$$

dado el dato X, se quiere predecir la secuencia de salida *Y*.

➢ Un **Modelo end-to-end** para RAH es un sistema que

- ✓ directamente hace corresponder una secuencia de entrada de características acústicas en una secuencia de grafemas o palabras.
- ✓ está entrenado para optimizar un criterio relacionado con la métrica de evaluación en la que estemos interesados (por ejemplo, WER).

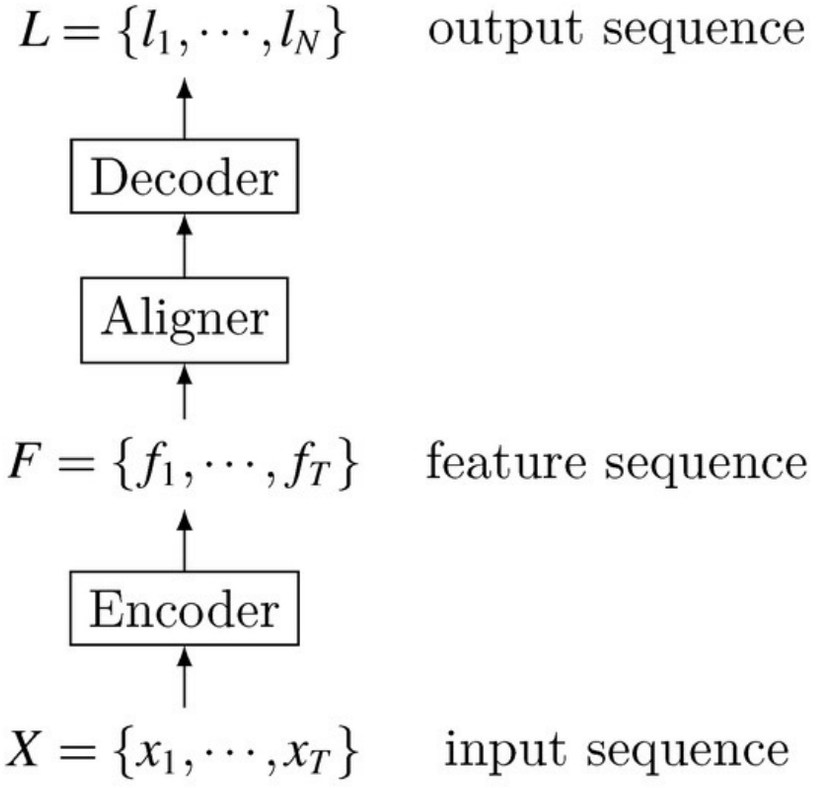
Escalado hacia End-to-end Speech Recognition

- Habitualmente, Deep Speech no requiere de un diccionario de fonemas, pero utiliza un sistema de entrenamiento RNN optimizado (múltiples GPU). Las GPU se usan porque el modelo se entrena utilizando miles de horas de voz.
- El principal componente de Deep Speech es una RNN entrenada para “ingerir” señal acústica (espectrogramas) y generar transcripciones de texto. El propósito de la RNN es convertir una secuencia de entrada (voz) en una secuencia de probabilidades de caracteres para la transcripción.

Modelo End-to-End

La mayoría de los sistemas end-to-end incluyen las siguientes partes:

- *encoder*, que mapea la secuencia de entrada de voz con la secuencia de características;
- *aligner*, que realiza el alineamiento entre la secuencia de características y el lenguaje;
- *decoder*, que decodifica el resultado identificado final.



Esta división no siempre existe, porque la arquitectura end-to-end es una estructura completa y, por lo general, es muy difícil saber qué parte realiza qué subtarea en la analogía de un sistema modular convencional.

From An Overview of End-to-End Automatic Speech Recognition (Symmetry 2019)

Modelo End-to-End

- Múltiples módulos se fusionan en una Red Profunda (*Deep Net)* para un entrenamiento conjunto (*joint-training*)
  - Ventaja: reemplaza múltiples módulos con una red profunda, de forma que no hay necesidad de diseñar muchos módulos para realizar el mapeo entre varios estados intermedios.
  - La función de optimización es global.
- El modelo *end-to-end* hace el mapeo directo de señales acústicas en secuencias de etiquetas sin estados intermedios cuidadosamente diseñados. Además, no hay necesidad de realizar un procesamiento posterior en la salida.

Tareas Sequence-to-Sequence

En distinto grado, las tareas sequence-to-sequence se enfrentan a problemas de alineamiento de datos, especialmente para reconocimiento de voz: **una etiqueta de la secuencia de etiquetas (caracteres o palabras) se debe alinear con señal acústica.**

- Modelado Sequence-to-Sequence:
  - *traducción*: secuencia de palabras (discreta) → secuencia de palabras (discreta)
  - *síntesis*: secuencia de palabras (discreta) → señal acústica (continua)
  - *reconocimiento del habla*: señal acústica (continua) → secuencia de palabras (discreta)
- Las longitudes de las secuencias pueden diferir en ambos lados:
  - señal *acústica muestreada a 10ms/5ms* - T-length $X_{1:T}$
  - *secuencias* de palabras/tokens - L-length $\omega_{1:L}$

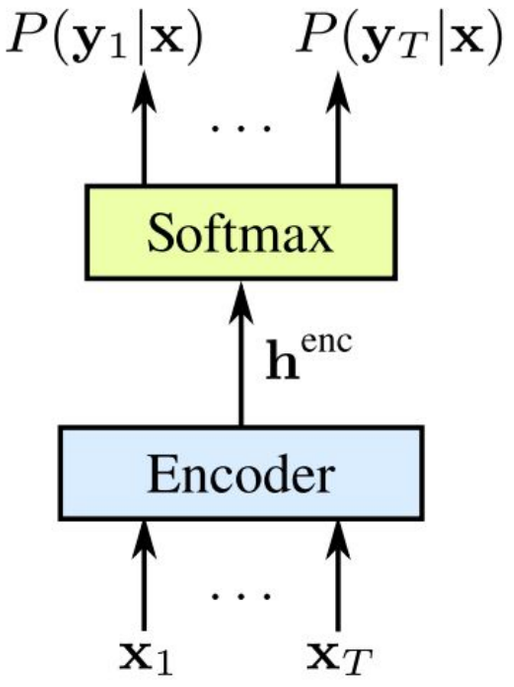
Tareas Sequence-to-Sequence

El modelo end-to-end utiliza **soft alignment** : cada frame de audio se corresponde con todos los posibles estados con una determinada distribución de probabilidad, que no requiere una correspondencia forzada y explícita.

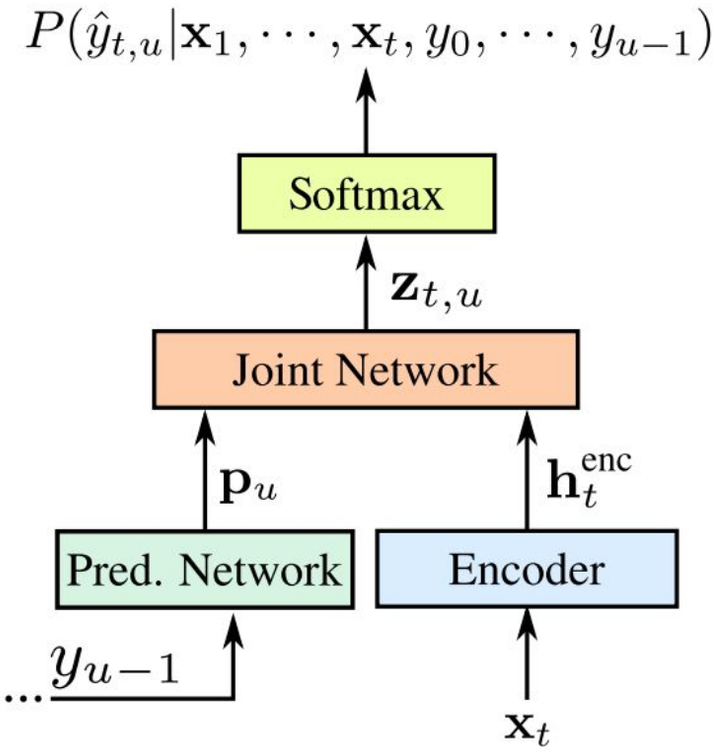
El modelo end-to-end se puede dividir en tres **categorías** según cómo se implemente el **soft alignment:**

- basado en modelos CTC (Connectionist Temporal Classification)
- transductor RNN
- basado en modelos de atención

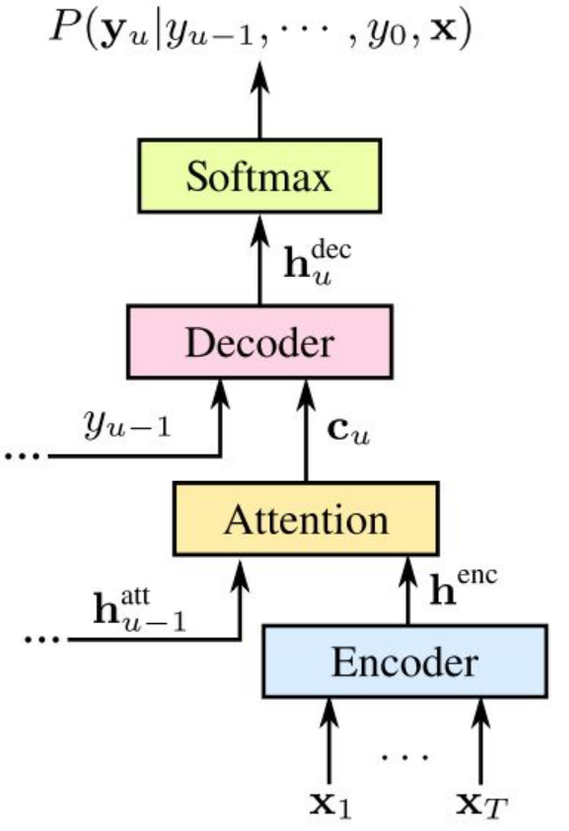
Tareas Sequence-to-Sequence



CTC-based

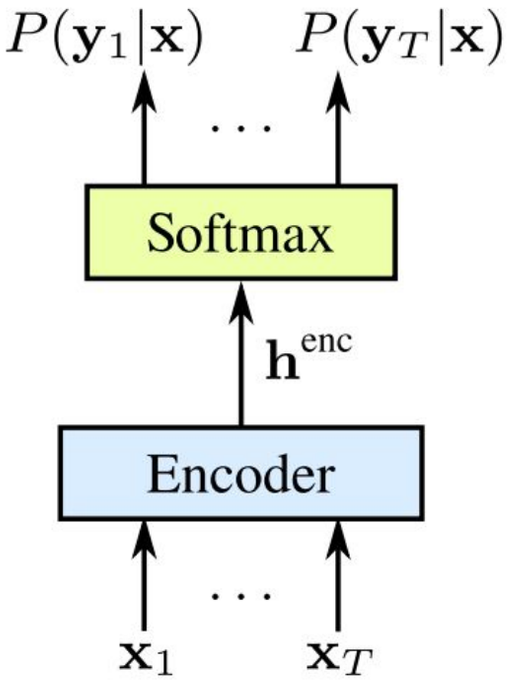


RNN-transducer



Attention-based

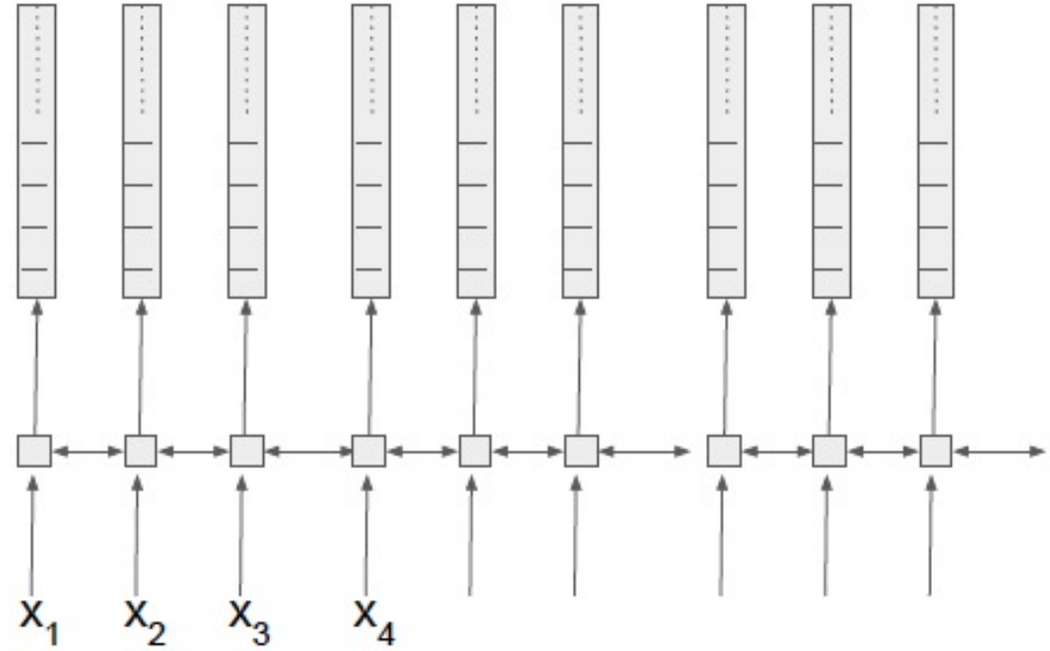
CTC-based



- CTC permite entrenar un modelo acústico sin necesidad de alineaciones a nivel de frame entre la señal acústica y las transcripciones.
- CTC primero enumera todas las posibles alineaciones y luego logra un **soft alignment** agregando todas estas alineaciones iniciales.
- CTC asume que las etiquetas de salida son independientes entre sí al enumerar todas las posibles alineaciones.

Encoder: fomado por múltiple capas de RNNs, unidireccionales o bidirecciones (la mayoría de las veces, LSTMs).

CTC model

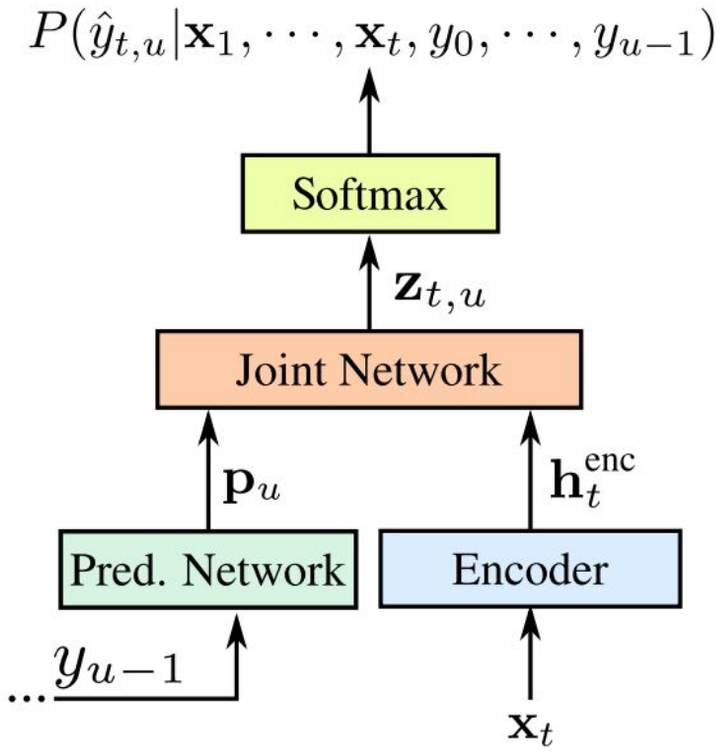


Softmax over vocabulary
{a, b, c, d, e, f, .. z, ?, ., !,..} and extra token <b>.

Softmax at step, t, gives a score s(k,t) = log Pr(k, t|**X**) to category k in the output at time t.

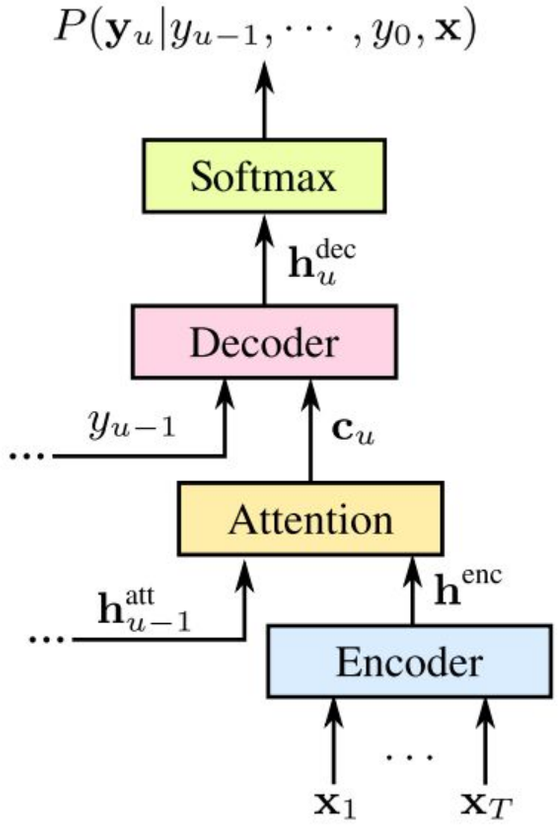
X es la secuencia de entrada de tramas acústicas: x1, x2, …, XT
Y es la secuencia de palabras de salida: w1, w2, …, wN

RNN-transducer



- El transductor RNN también enumera todas las posibles alineaciones y luego las agrega para una **soft alignment**.
- Pero a diferencia de CTC, el transductor RNN no hace suposiciones de independencia sobre las etiquetas al enumerar las alineaciones, por lo que es diferente de CTC en términos de cálculo de probabilidad.

Attention-based



- Los métodos basados en modelos de atención ya no enumeran todas las posibles alineaciones, sino que utilizan el mecanismo de atención para calcular directamente la **soft alignment** entre los datos de entrada y la etiqueta de salida.
- Partes:
  - **Encoder** (análogo al modelado acústico): transforma la entrada de voz en una representación superior
  - **Attention** (modelo de alineamiento): identifica las frames codificadas que son relevantes para generar una salida
  - **Decoder** (análogo al modelo de lenguaje y de pronunciación): funciona prediciendo cada salida como una función de las predicciones previas

Mecanismo de atención: resume las características del codificador relevantes para predecir la siguiente etiqueta

Transformer models

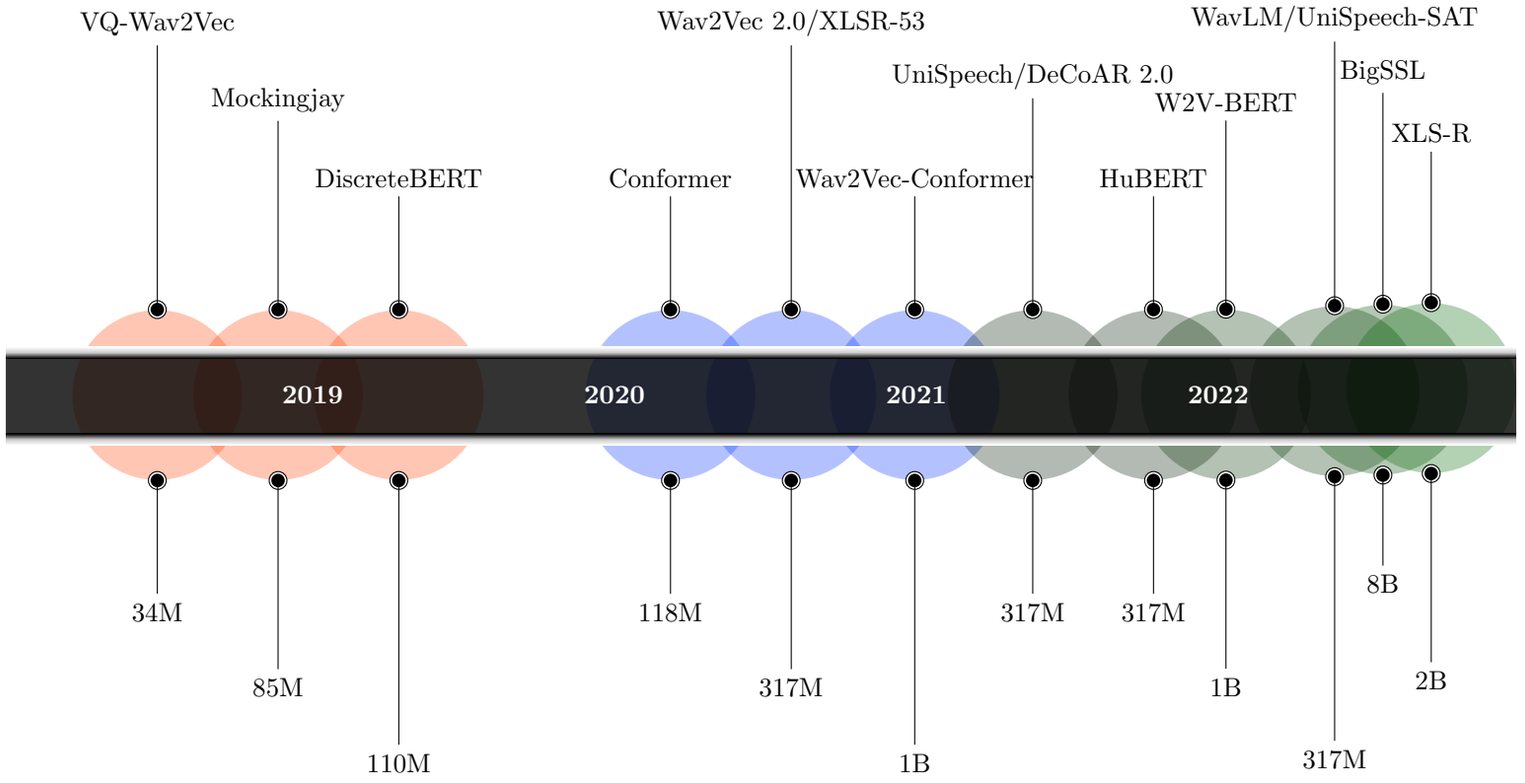
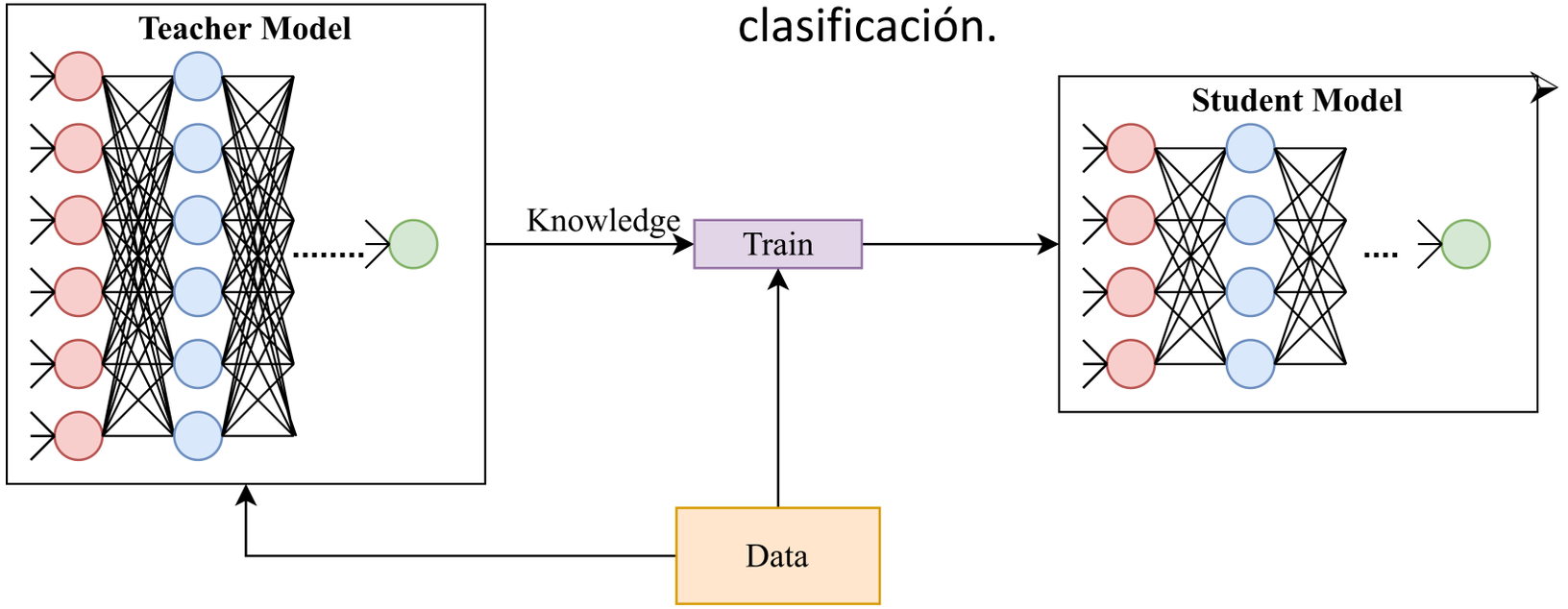


Fig. 4. Timeline highlighting notable large Transformer models developed for speech processing, along with their corresponding parameter sizes.

Knowledge Distillation

- *Knowledge distillation* es una técnica de aprendizaje automático que busca reducir el tamaño de un modelo de RN, mientras se mantiene o mejora su precision.
- Esto se logra mediante el proceso de "destilación" (transferencia de conocimiento) de un modelo grande y complejo (llamado "maestro") a uno más pequeño y sencillo (llamado "estudiante").
- El proceso se basa en entrenar el modelo estudiante para que imite las decisiones del modelo maestro. El modelo maestro se utiliza como una guía para enseñar al modelo estudiante, ya que tiene una mayor capacidad y precisión. La idea es que el modelo estudiante aprenda las características importantes del modelo maestro y las aplique en su propio proceso de clasificación.



Si bien los modelos grandes (como las redes neuronales muy profundas) tienen una mayor capacidad de conocimiento que los modelos pequeños, es posible que esta capacidad no se utilice por completo y es difícil implementar dichos modelos en dispositivos con recursos limitados, como teléfonos móviles.

¿RAH en el año 2030?

- Aprendizaje semi-supervisado. En particular, los modelos preentrenados auto-supervisados serán parte de muchas aplicaciones de aprendizaje automático, incluido RAH.
- La mayor parte del RAH se hará en el propio dispositivo.
- Los investigadores ya no publicarán artículos del tipo “WER mejorada con corpus de referencia X usando el modelo Y” ¡WER saturado!
- Las transcripciones serán reemplazadas por representaciones enriquecidas para tareas que dependen de la salida de un reconocedor de voz. Ejemplos de aplicaciones posteriores incluyen agentes conversacionales, consultas de búsqueda basadas en voz y asistentes digitales.
- Los modelos de reconocimiento de voz estarán muy personalizados para usuarios individuales.

The History of Speech Recognition to the Year 2030 – Awni Hannun (2021)

¿Aplicaciones RAH en el año 2030?

- El 99% de los servicios de voz transcritos se realizarán mediante reconocimiento automático. Los transcriptores humanos realizarán un control de calidad y corregirán o transcribirán las expresiones más difíciles. Los servicios de transcripción incluyen, por ejemplo, subtitular videos, transcribir entrevistas y transcribir conferencias o discursos.
- Los asistentes de voz mejorarán, pero no de forma sustancial. El reconocimiento de voz ya no es el cuello de botella para mejores asistentes de voz. Los cuellos de botella ahora serán la comprensión, sistemas conversacionales, respuestas contextuales, y preguntas y respuestas de dominio mucho más amplio.
- ¿Viviremos en hogares inteligentes que siempre están escuchando y pueden responder a cada una de nuestros deseados? No. ¿Llevamos gafas de realidad aumentada en la cara y las controlamos con la voz? No para 2030.

The History of Speech Recognition to the Year 2030 – Awni Hannun (2021)