Transcrição automática de áudio em texto com Deep Learning

Olá!

Arthur Fortes

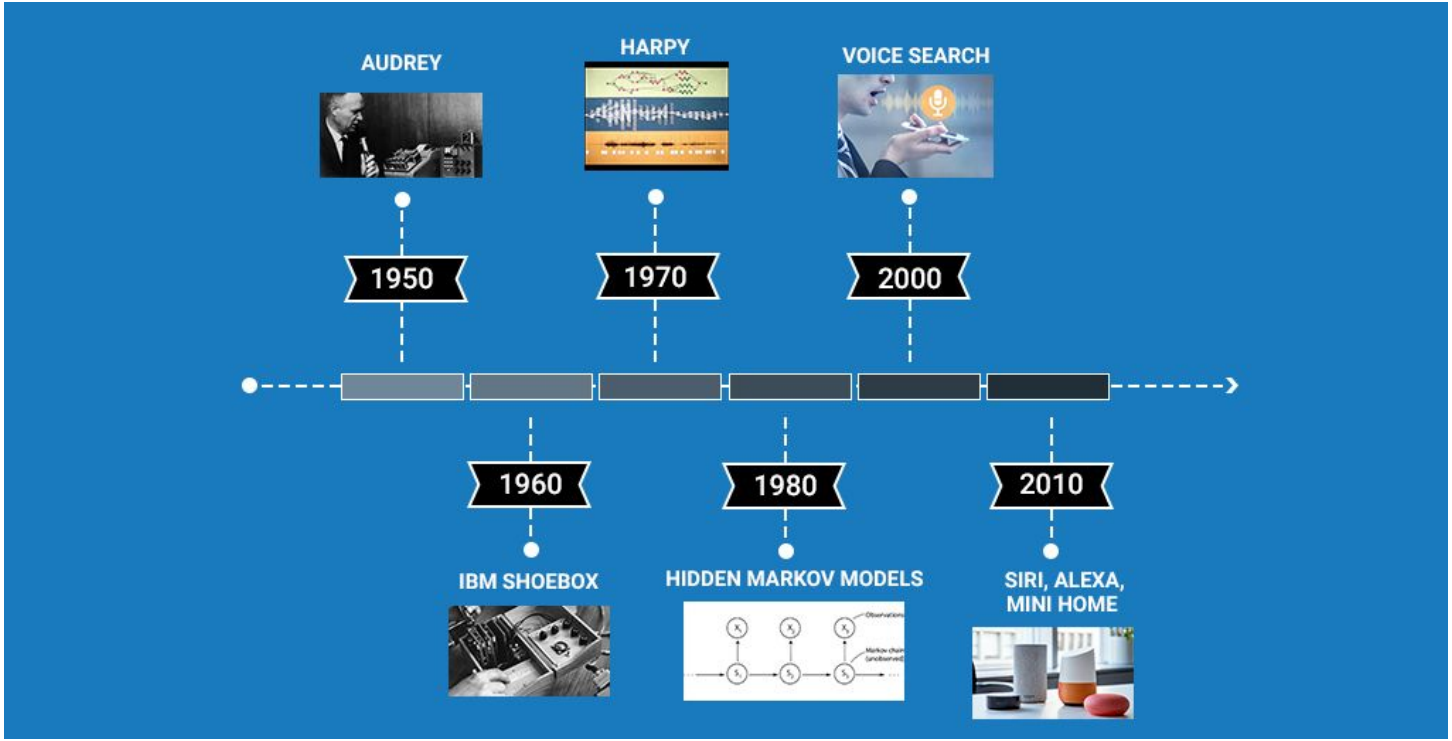
Cientista de Dados @ Instituto de Pesquisas Eldorado
Doutor em Computação @ ICMC - USP

arthurfortes.github.io

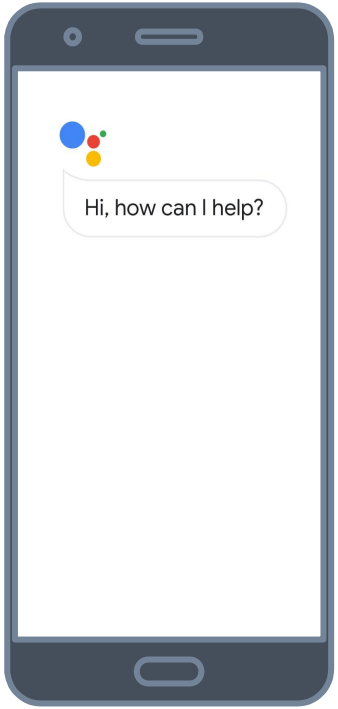
Agenda

1. Uma Breve História do Reconhecimento de Fala
2. Entendendo como funciona
3. Processamento de Sinais
4. Reconhecendo caracteres de sons curtos
5. Demonstração

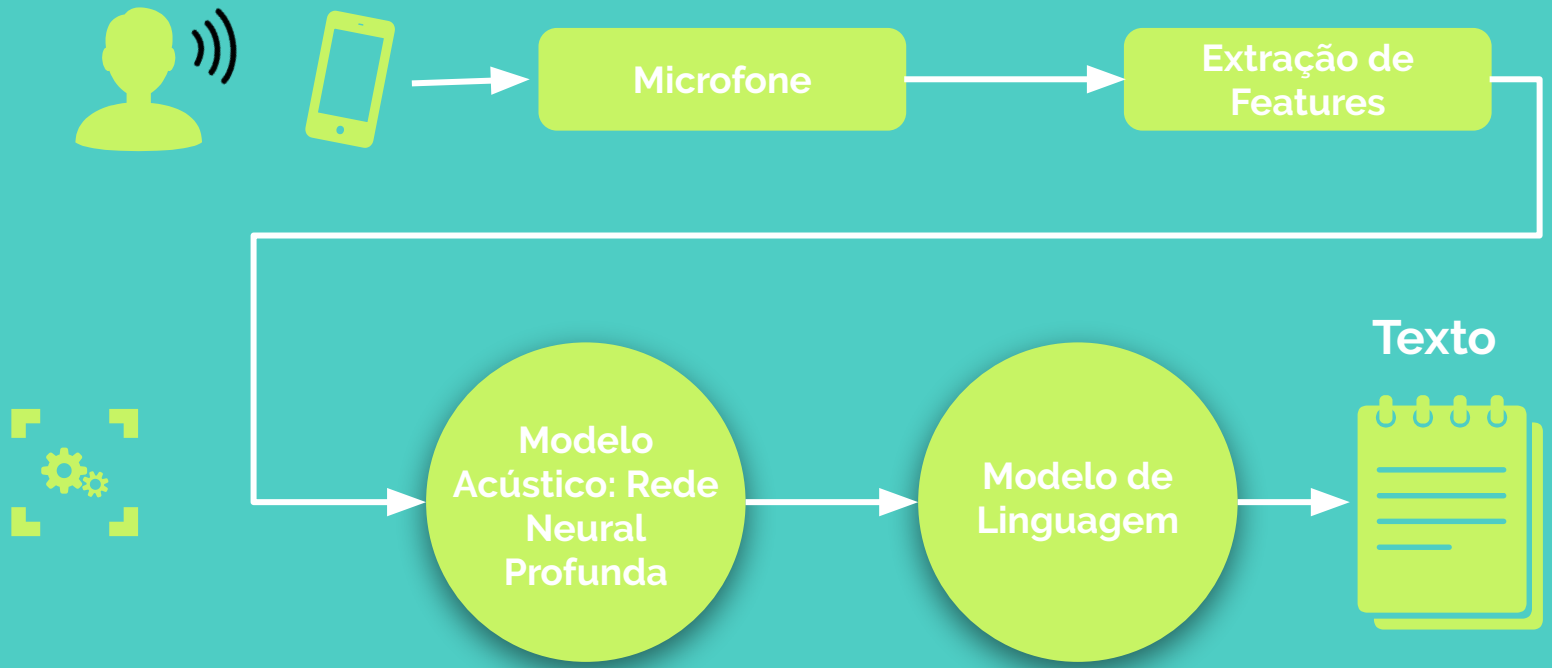
1950 - 2010


AUDREY
HARPY
VOICE SEARCH
1950
1970
2000
1960
1980
2010
IBM SHOEBOX
HIDDEN MARKOV MODELS
SIRI, ALEXA, MINI HOME

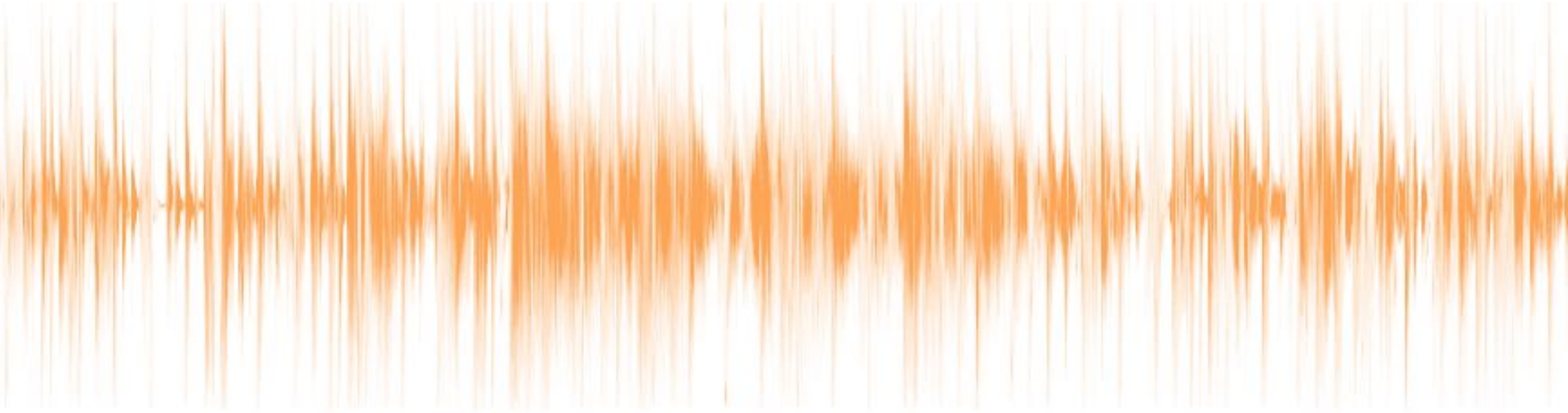
2019



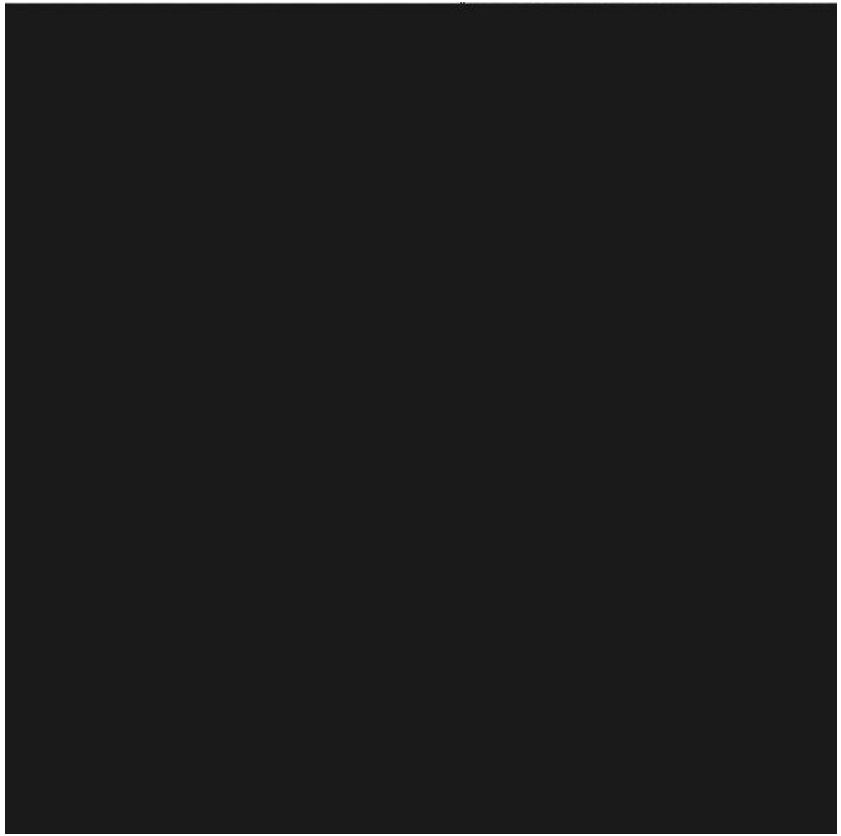
"Alexa, play morning playlist."



Como funciona?



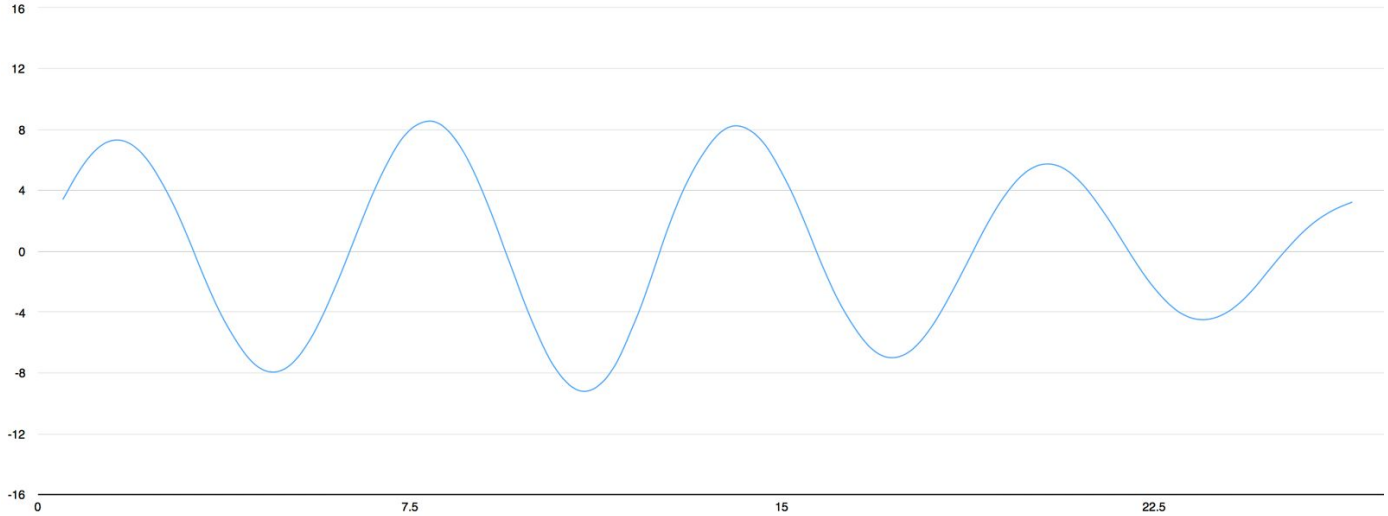
Processamento de sinais



Em classificação de imagens, transforma-se o arquivo em uma matriz numérica.

Processamento de Sinais

Para transformar uma onda sonora em números, uma das maneiras é registramos a altura da onda em pontos igualmente espaçados:



Processamento de Sinais

- Estamos fazendo uma leitura milhares de vezes por segundo e gravando um número que representa a altura da onda sonora naquele momento. Basicamente, trata-se de um arquivo de áudio .wav não compactado.

- O áudio "Qualidade do CD" é amostrado em 44,1 kHz (44.100 leituras por segundo). Mas para o reconhecimento de fala, uma taxa de amostragem de 16khz (16.000 amostras por segundo) é suficiente para cobrir a faixa de frequência da fala humana.

Processamento de Sinais

Cada número representa a amplitude da onda sonora em intervalos de 1/16000 de segundo

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

Graças ao teorema de Nyquist, sabemos que podemos usar a matemática para reconstruir perfeitamente a onda sonora original a partir de amostras espaçadas - desde que amostremos pelo menos duas vezes mais rápido que a frequência mais alta que queremos gravar.

Processamento de Sinais

Em Python:

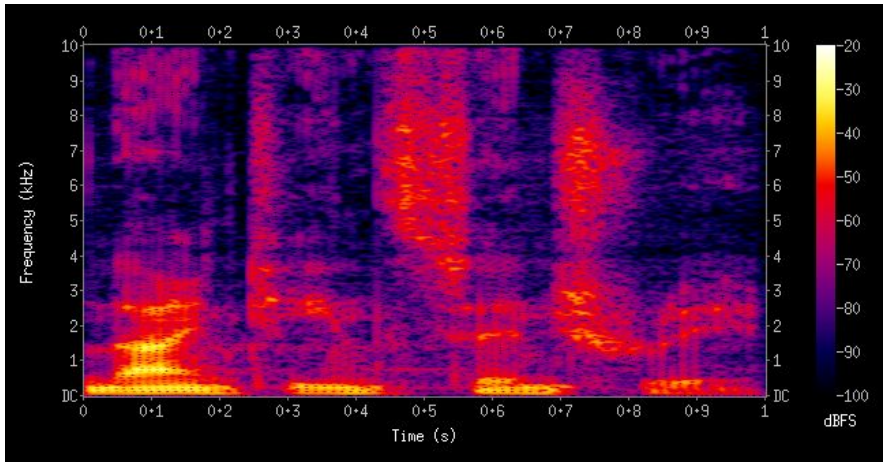
```
from scipy.io import wavfile

fs, data = wavfile.read('audio.wav')
```

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.io.wavfile.read.html

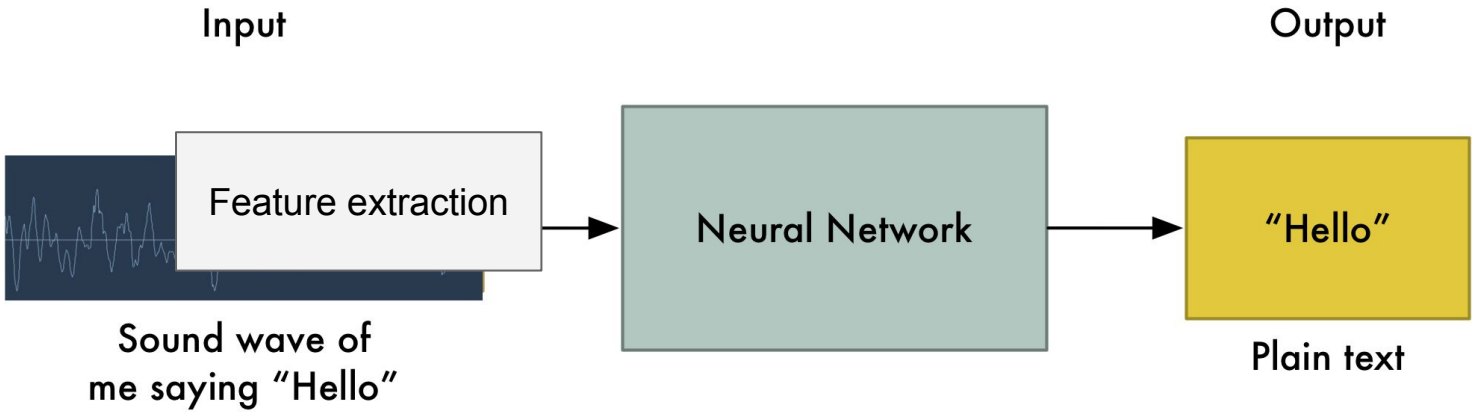
Processamento de Sinais

Nós podemos também representar o áudio através de um espectrograma, que é um gráfico 2D entre tempo e frequência, em que cada ponto no gráfico representa a amplitude de uma frequência específica em um momento específico em termos de intensidade de cor.



Em termos simples, o espectrograma é um espectro (ampla gama de cores) de frequências, pois varia com o tempo.

Cenário ideal



Problemas?

- Vocabulário
- Velocidade da fala
- Gírias
- Sotaques

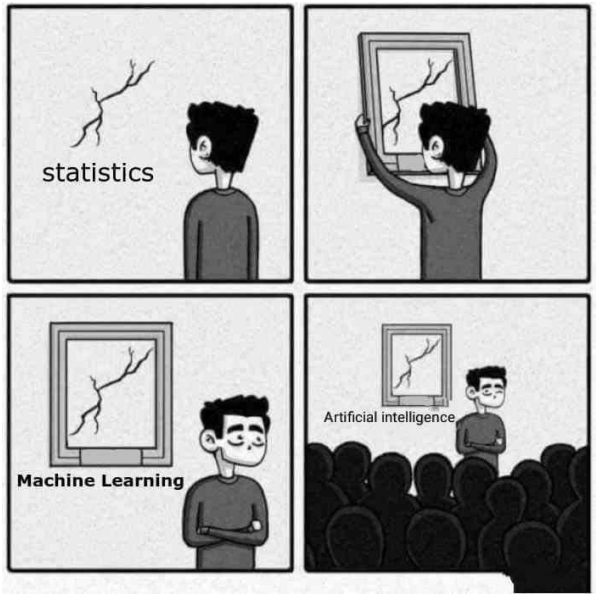
Deep Learning

Reconhecendo caracteres de sons curtos

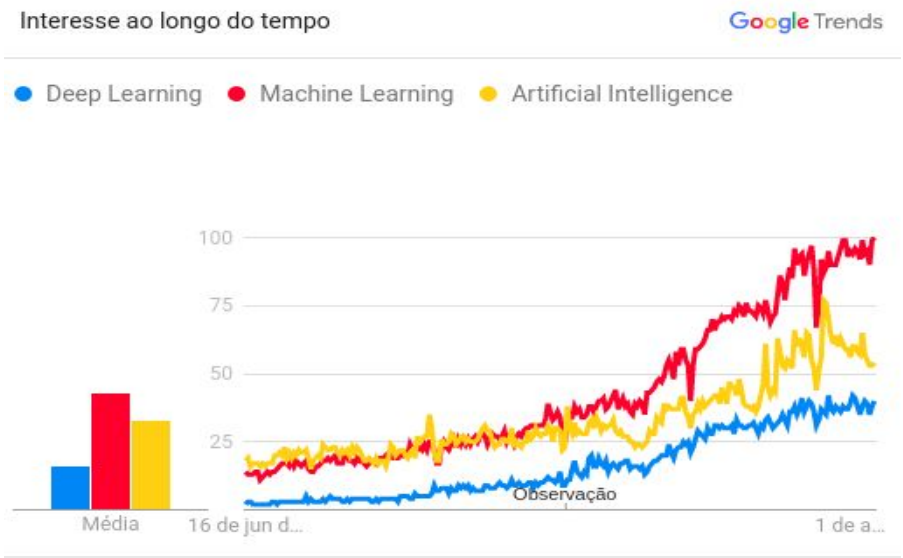
Mas o que é e como funciona deep learning?

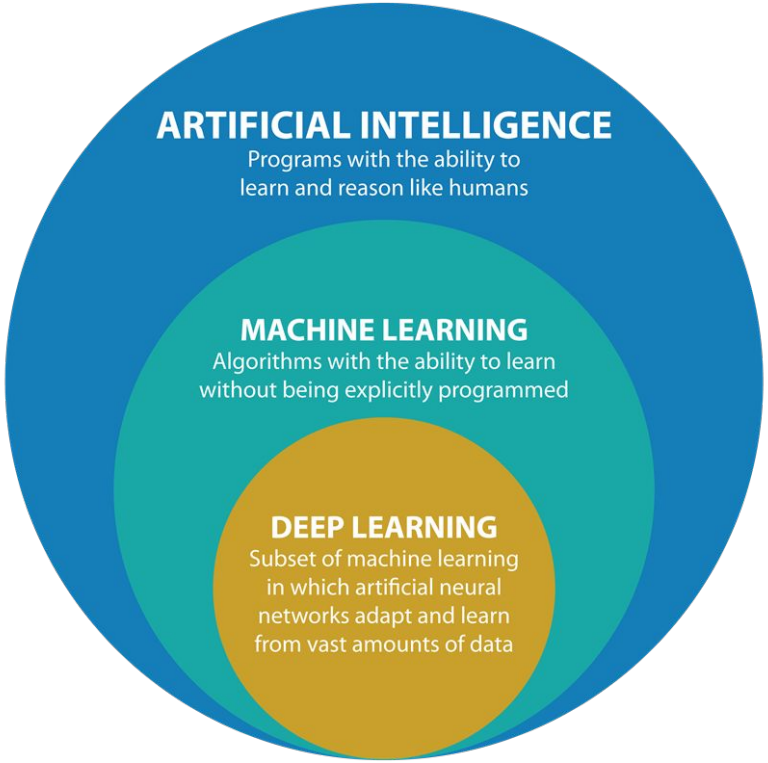
Inteligência Artificial, Machine Learning e Deep Learning

São três termos que muitas vezes são confundidos entre si. Antes de falar exclusivamente sobre Deep Learning, é importante contextualizar a hierarquia dessa área de estudo e as diferenças entre os termos, que muitas vezes são utilizados de forma indiscriminada e generalizada.



Inteligência Artificial, Machine Learning e Deep Learning





Inteligência Artificial, Machine Learning e Deep Learning

- **Inteligência Artificial (IA)** é a inteligência similar à humana exibida por mecanismos ou software (agentes inteligentes).
- **Machine Learning (ML)** é um subconjunto de técnicas, que de forma geral, “aprendem” a tomar uma decisão baseadas em exemplos de um problema, e não de uma programação específica. (Necessitam de dados)
- Um subgrupo específico de técnicas de ML são chamadas de **Deep Learning (DL),** geralmente utilizam redes neurais profundas e dependem de muitos dados para o treinamento.

Machine Learning x Deep Learning

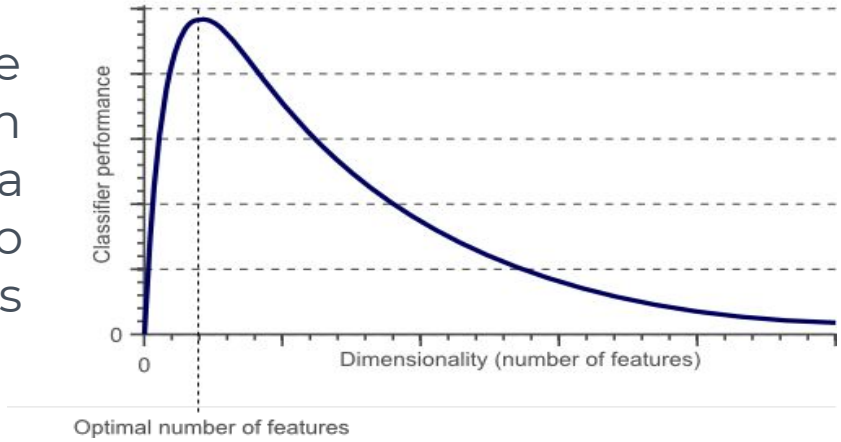
- Existem alguns pontos importantes que diferem as técnicas clássicas de ML dos métodos de DL, os principais são: a **necessidade e o efeito de muitos dados, poder computacional** e a **flexibilidade na modelagem dos problemas**.

- **Segredo:**

“Não existe Machine Learning se não existir dados.”

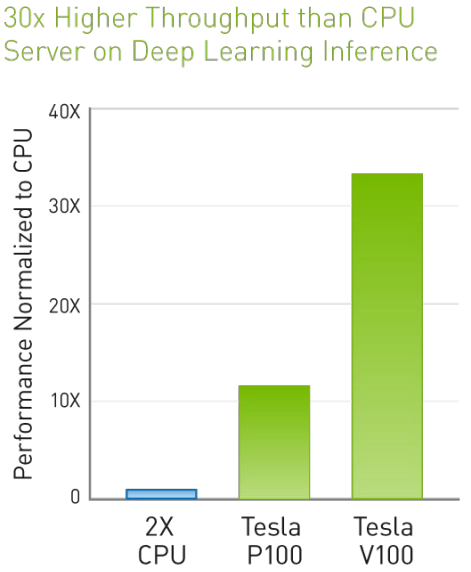
Machine Learning x Deep Learning

- Existem dois problemas que afligem as técnicas clássicas com relação aos dados, a “maldição da dimensionalidade” e a estagnação da performance ao adicionar mais dados após um certo limite.



Machine Learning x Deep Learning

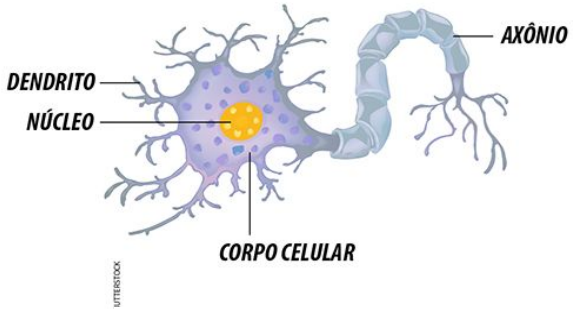
- Técnicas de DL são preparadas para trabalhar com maior quantidade de dados. E possuem um maior poder computacional.

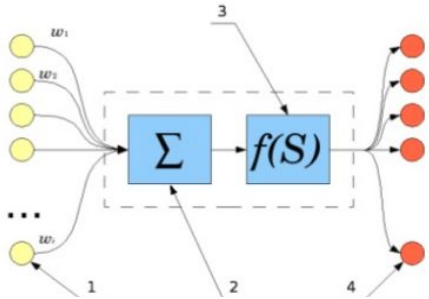


Workload: ResNet-50 | CPU: 2X Xeon E5-2660 v4, 2GHz | GPU: add 1X Tesla P100 or V100 at 150W | V100 measured on pre-production hardware.

Deep Learning

- Embora existam diversas técnicas clássicas e sejam de propósito geral, a estrutura de Deep Learning e sua unidade mais básica, o neurônio, consegue ser o mais genérico e flexível possível.
- A formulação matemática de um neurônio (esse mesmo que tem ai na sua cabeça) é o mais simples possível, e fazendo um paralelo nós temos que:



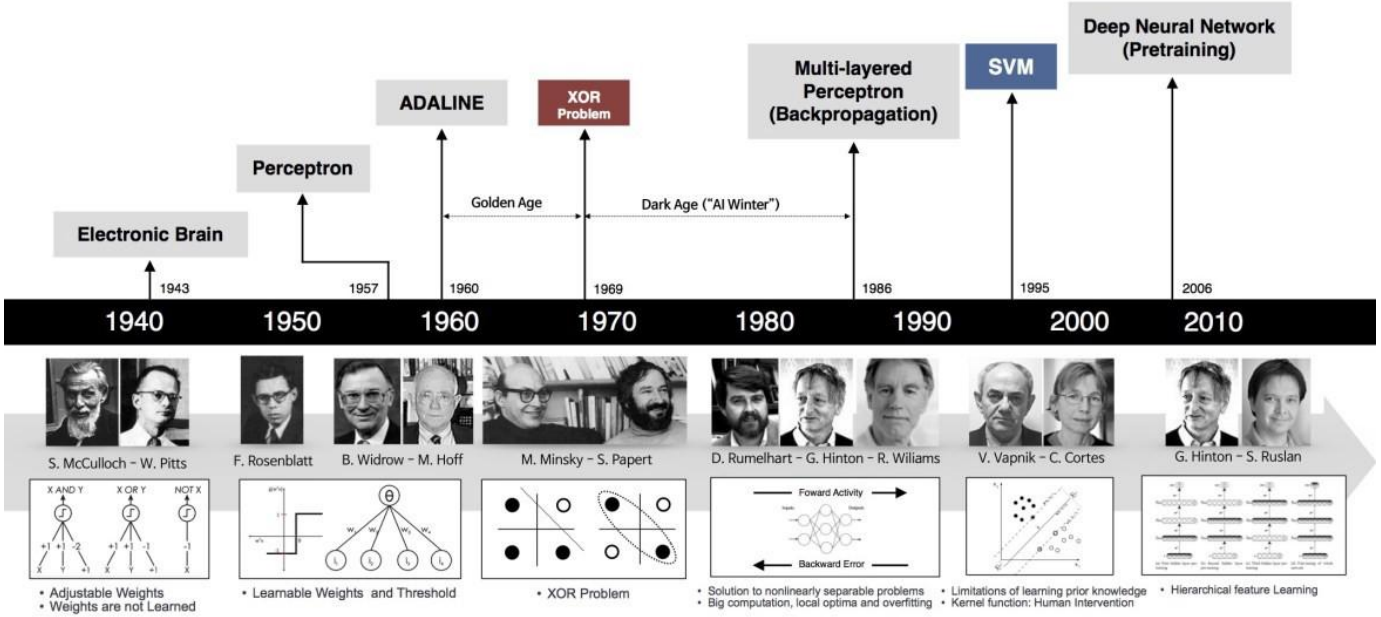


$$S = \sum_{i=1}^{n} x_i.w_i + bias$$

$$out = f(S)$$

Baseado em “A Logical Calculus of the ideas Immanent in Nervous Activity” McCulloch and Pitts, 1943

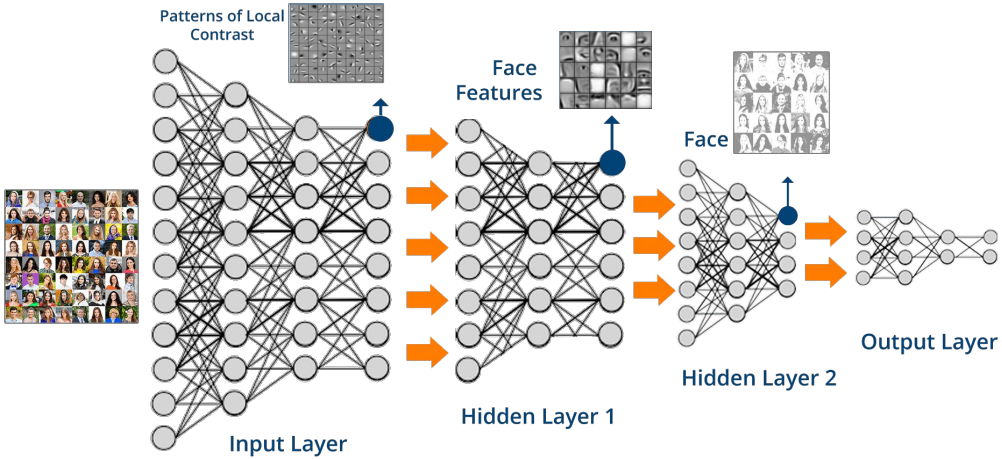
Deep Learning



Uma breve história das redes neurais artificiais — http://deeplearningbook.com.br/uma-breve-historia-das-redes-neurais-artificiais/

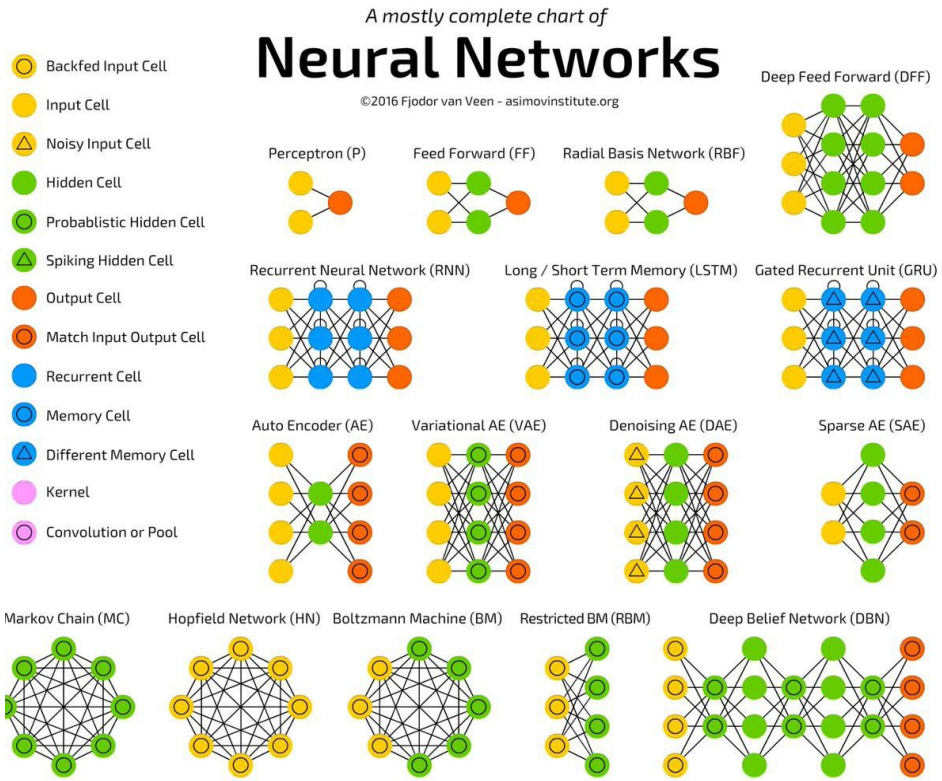
Como funciona Deep Learning?

- Abaixo tem uma estrutura simplificada de uma Rede Neural Convolucional (CNN). Esse tipo de rede é muito utilizada para reconhecimento facial, detecção de objetos em imagens, extração de características.. etc.



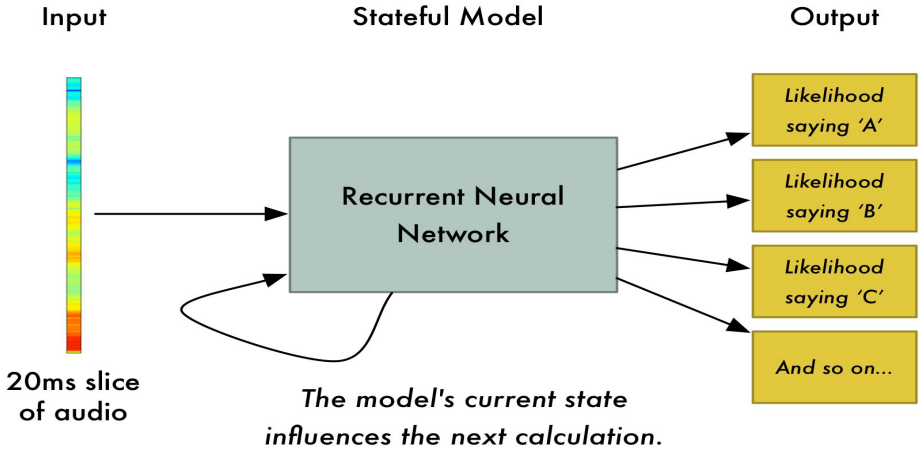
Como funciona Deep Learning?

- Aplicações:
  - Visão computacional
  - Reconhecimento de fala
  - Healthcare
  - Sistemas de Recomendação
  - Detecção de fraudes
  - Análise de sentimento



Reconhecendo caracteres

- Geralmente usamos chunks de audio de 20 milisegundos como entrada das redes.



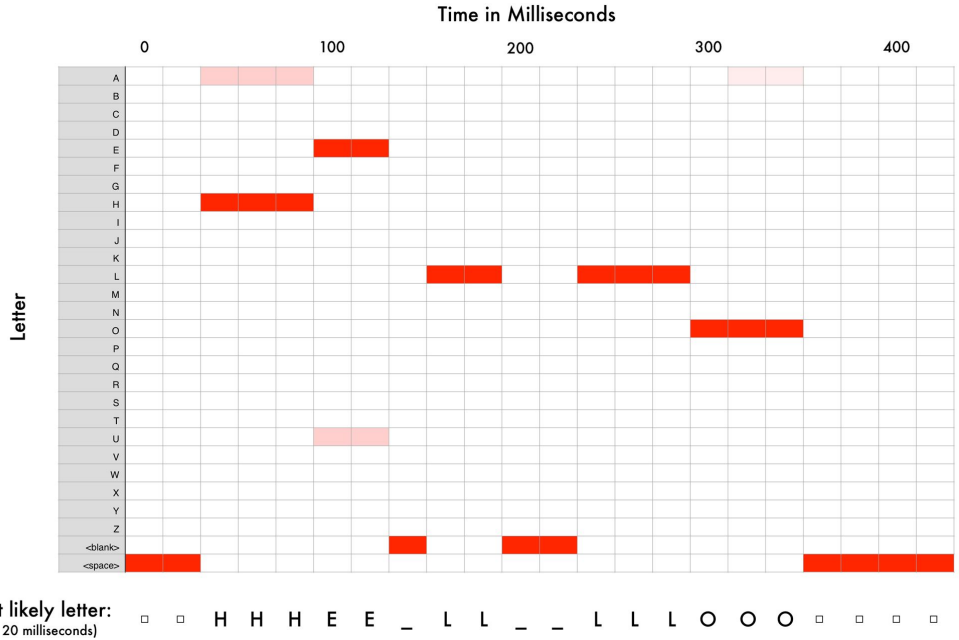
- Para cada fatia de áudio, ele tentará descobrir a letra que corresponde ao som que está sendo falado atualmente:

Reconhecendo caracteres

- Rede neural recorrente: uma rede neural que possui uma memória que influencia previsões futuras.
- Isso porque cada letra que ela prevê deve afetar a probabilidade da próxima carta que ela também prevê.
- Por exemplo, se dissermos "HEL" até agora, é muito provável que digamos "LO" ao lado para finalizar a palavra "HELLO".

Reconhecendo caracteres

- A saída da rede neural é um mapeamento de cada pedaço de áudio para as letras provavelmente faladas durante esse pedaço.



- Possibilidades:
  1. “**HHHEE_LL_LLLOOO**”
  2. “HHHUU_LL_LLLOOO”
  3. “AAAUU_LL_LLLOOO”

Reconhecendo caracteres

- Temos alguns passos que seguimos para limpar essa saída. Primeiro, substituiremos qualquer caractere repetido por um único caractere:
  - *HHHEE_LL_LLLOOO -> HE_L_LO*
  - *HHHUU_LL_LLLOOO ->HU_L_LO*
  - *AAAUU_LL_LLLOOO -> AU_L_LO*
- Em seguida, removeremos todos os espaços em branco (_):
  - *HE_L_LO -> HELLO*
  - *HU_L_LO -> HULLO*
  - *AU_L_LO -> AULLO*

Reconhecendo caracteres

De nossas possíveis transcrições “Hello”, “Hullo” e “Aullo”, obviamente “Hello” aparecerá com mais frequência em um banco de dados de texto (sem mencionar em nossos dados originais de treinamento baseados em áudio) e, portanto, provavelmente está correto.

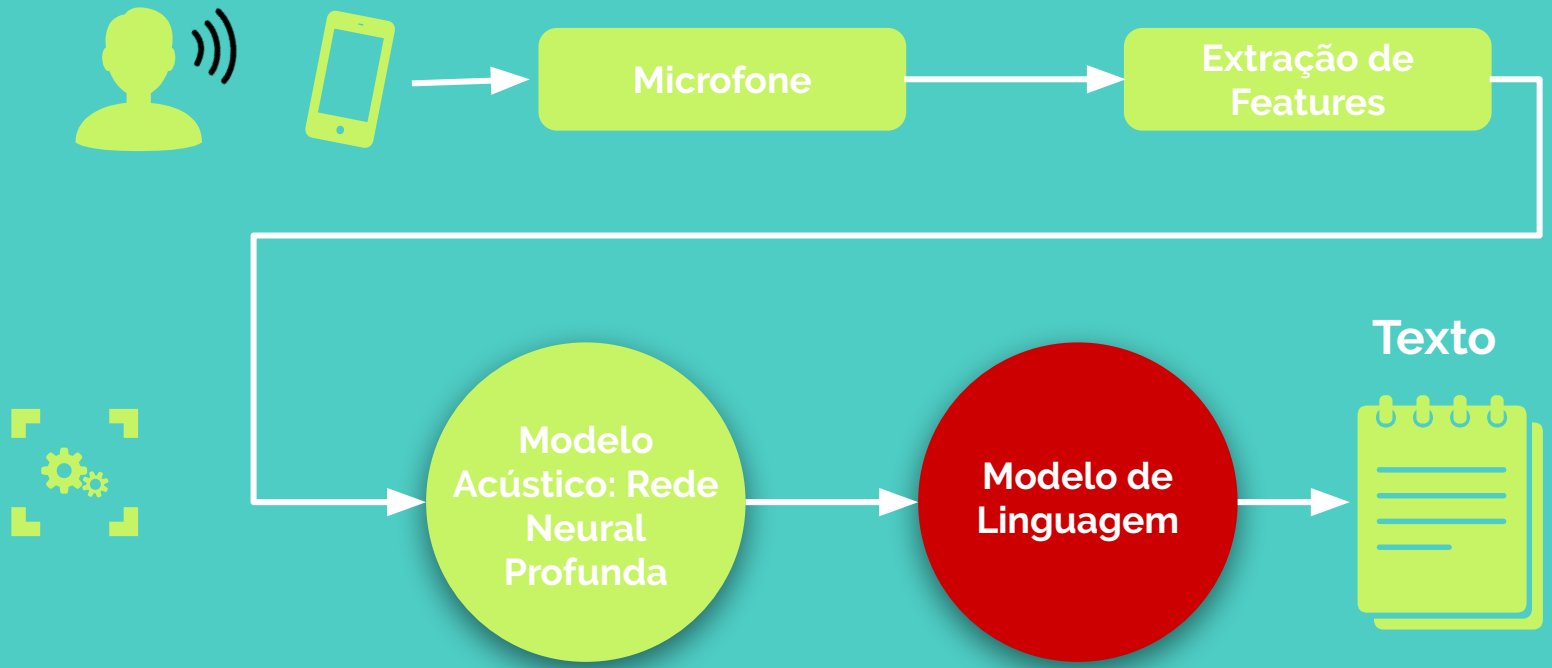
Mas calma. E se alguém realmente disser “Hullo”?

Excessões

- Não reconhecer "Hullo" é um comportamento razoável, mas às vezes você encontrará casos irritantes em que seu telefone simplesmente se recusa a entender algo válido que você está dizendo.

- É por isso que esses modelos de reconhecimento de fala estão sempre sendo treinados novamente com mais dados para corrigir esses casos de borda.

- Modelos de linguagem também são responsáveis por auxiliar nessa função, pois servem como um corretor para esses modelos.



Como funciona?

Demonstração

Obrigado!

Perguntas?

@fortesarthur | @arthurfortes
fortes.arthur@gmail.com