

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Licenciatura en Ciencia de Datos

Introducción al Aprendizaje Profundo
Estrategias de entrenamiento

Profesores:
Berenice & Ricardo Montalvo Lezama

Mayo 2021

Contenido basado en el curso de AP del Dr. Gibran Fuentes Pineda del PCIC

Estrategias de entrenamiento

- Preprocesamiento.
- Aumentado de datos.
- Aceleración de convergencia.
- Regularización.
- Transferencia de conocimiento.

Preprocesamiento de datos

Normalización

- Estandarización:

$$x' = \frac{x - \bar{x}}{\sigma}$$

- Escalado:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Magnitud unitaria:

$$x' = \frac{x}{\|x\|}$$

Preprocesamiento de imágenes (I)

- Radio uniforme.

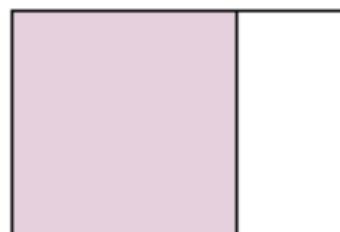
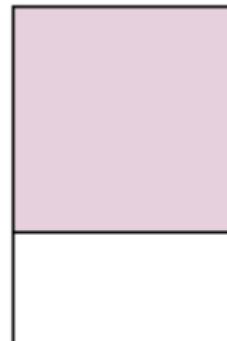
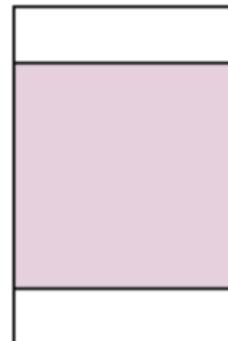
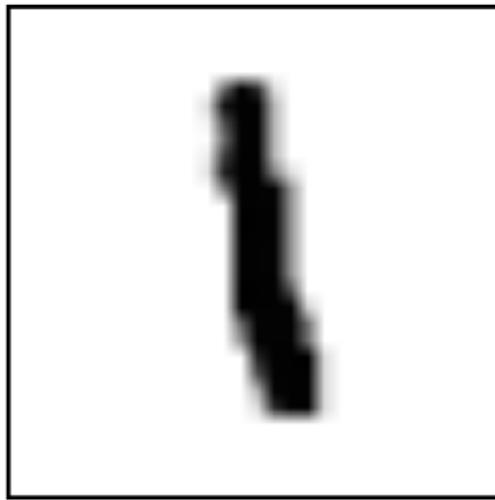


Imagen tomada de Nikhil B. Image Data Pre-Processing for Neural Networks, 2017.

Preprocesamiento de imágenes (II)

- Escalado de valores de pixeles.



|?

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	.6	.8	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	.7	.1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	.7	.1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	.5	.1	.4	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.1	.4	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.1	.4	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.1	.7	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.1	.1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.9	.1	.1	0	0	0	0	0	0
0	0	0	0	0	0	0	.3	.1	.1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Imagen tomada de presentación de Yubei Chen. Part I: Manifold Learning, 2018.

Preprocesamiento de audio (I)

- Filtrado de bandas.

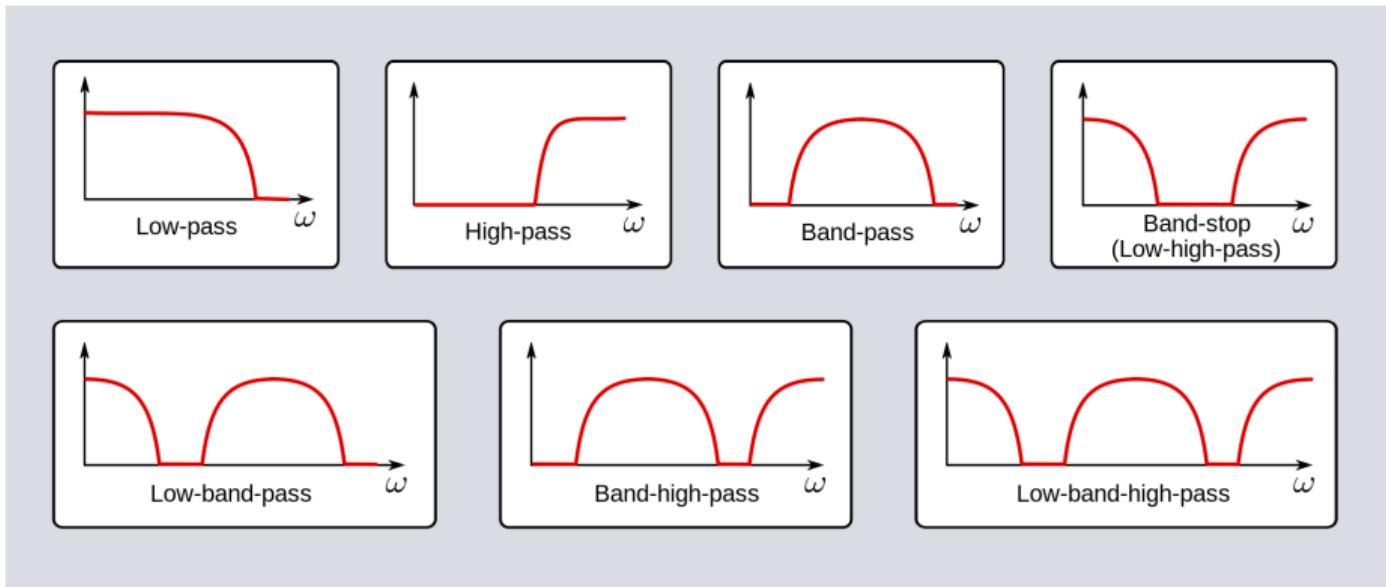


Imagen del usuario SpinningSpark de Wikipedia (entrada Filter (signal processing)). CC BY-SA 3.0

Preprocesamiento de audio (II)

- Espectogramas.

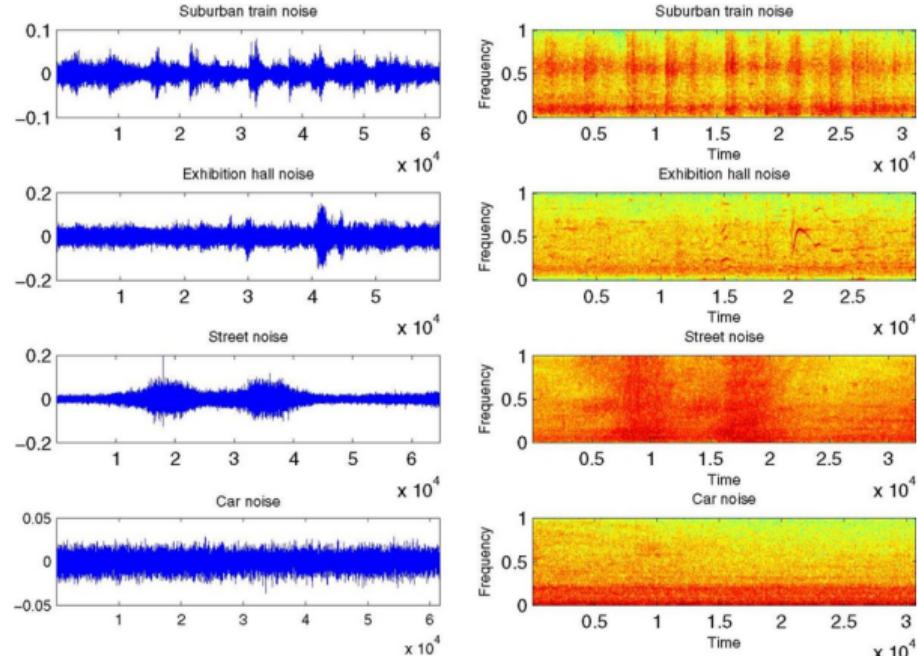


Imagen tomada de Zouhir and Ouni. A bio-inspired feature extraction for robust speech recognition, 2014.

Preprocesamiento de texto (I)

- Representaciones de bolsas de palabras.
 - Remover palabras demasiado comunes (stop words).
 - Reducir las palabras a sus raíces (stemming).
 - Quitar caracteres.



Imagen tomada de

https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

Preprocesamiento de texto (II)

- Representaciones distribuidas: encajes de palabra.

Source Text	Training Samples
The quick brown fox jumps over the lazy dog. ➔	(the, quick) (the, brown)
The quick brown fox jumps over the lazy dog. ➔	(quick, the) (quick, brown) (quick, fox)
The quick brown fox jumps over the lazy dog. ➔	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown fox jumps over the lazy dog. ➔	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

Imagen tomada de McCormick. Word2Vec Tutorial – The Skip-Gram Model, 2016.

Preprocesamiento de texto (III)

- Representaciones densas: word2vec (CBOW y skip-gram).

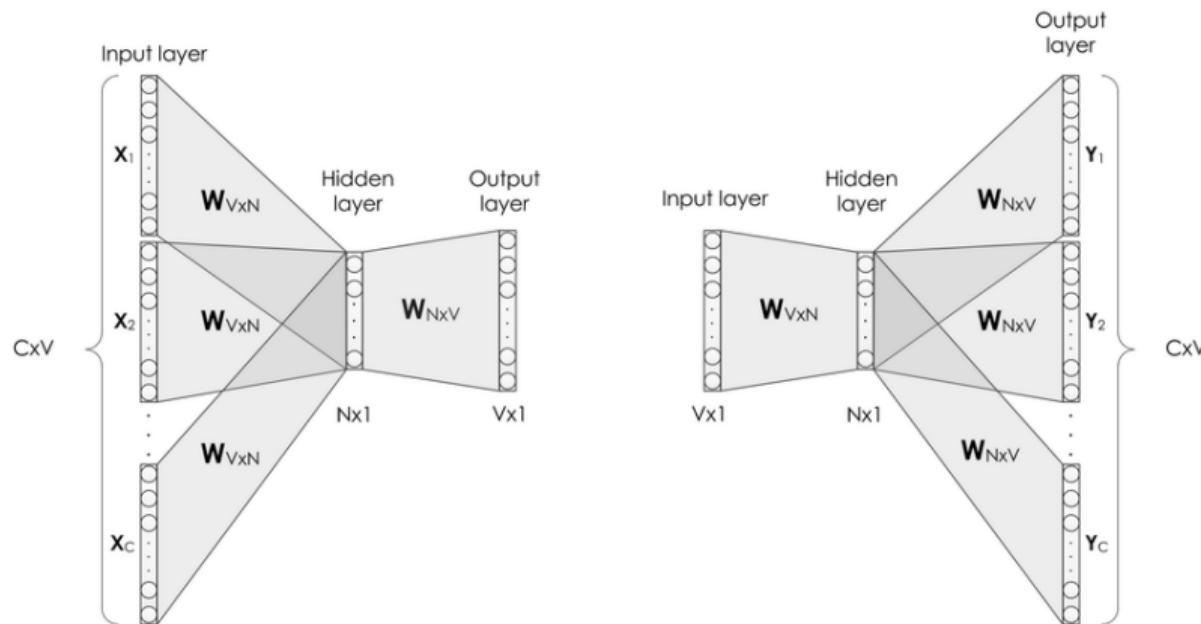


Imagen tomada de Tutubalina y Nikolenko. Demographic Prediction Based on User Reviews about Medications, 2017.

Preprocesamiento de video

- Preprocesamiento por marco.
- Muestreo de marcos (por ej., uniforme o por movimiento).
- Flujo de movimiento, segmentación de movimiento.



Imagen tomada de <https://www.commonlounge.com/discussion/1c2eaa85265f47a3a0a8ff1ac5fbce51>

Acrecentamiento de datos

Acrecentamiento de imágenes: transformaciones

- Traslación

$$\begin{bmatrix} 1 & 0 & 0 & v_x \\ 0 & 1 & 0 & v_y \\ 0 & 0 & 1 & v_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

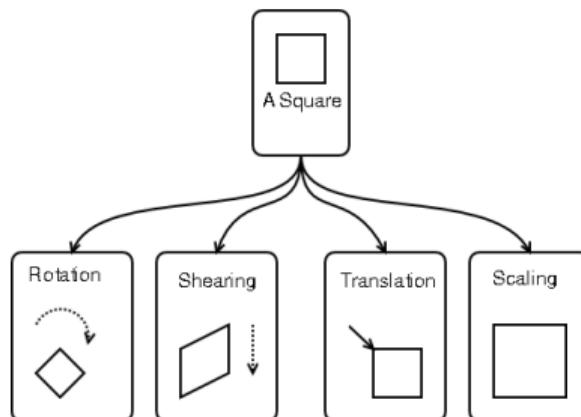


Imagen tomada de <https://people.gnome.org/~mathieu/libart/libart-affine-transformation-matrices.html>

Acrecentamiento de imágenes: transformaciones

- Traslación
- Rescalado

$$\begin{bmatrix} v_x & 0 & 0 \\ 0 & v_y & 0 \\ 0 & 0 & v_z \end{bmatrix}$$

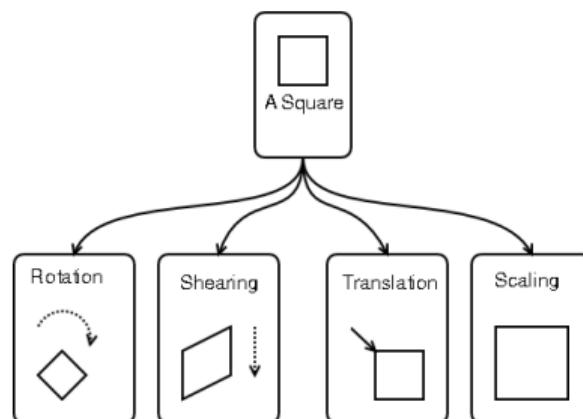


Imagen tomada de <https://people.gnome.org/~mathieu/libart/libart-affine-transformation-matrices.html>

Acrecentamiento de imágenes: transformaciones

- Traslación
- Rescalado
- Giro

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

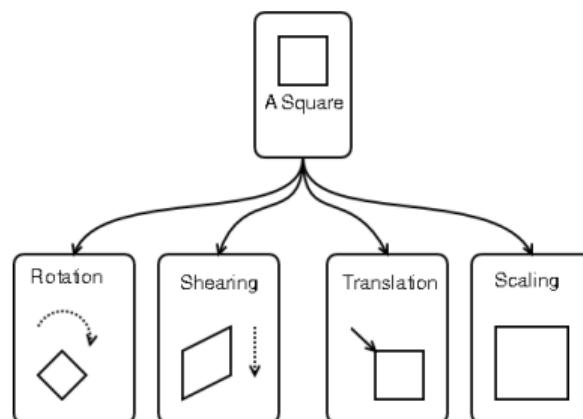


Imagen tomada de <https://people.gnome.org/~mathieu/libart/libart-affine-transformation-matrices.html>

Acrecentamiento de imágenes: transformaciones

- Traslación
- Rescalado
- Giro
- Rotación

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

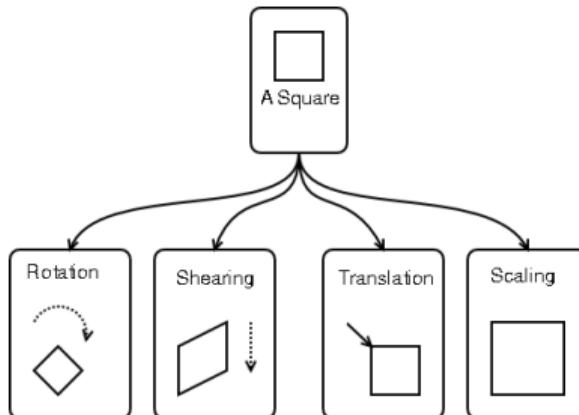


Imagen tomada de <https://people.gnome.org/~mathieu/libart/libart-affine-transformation-matrices.html>

Acrecentamiento de imágenes: transformaciones

- Traslación
- Rescalado
- Giro
- Rotación
- Deformación de corte

$$\begin{bmatrix} 1 & c_x = 0.5 & 0 \\ c_y = 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

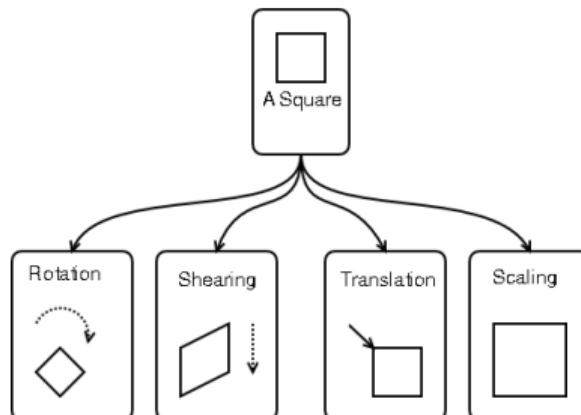


Imagen tomada de <https://people.gnome.org/~mathieu/libart/libart-affine-transformation-matrices.html>

Acrecentamiento de imágenes: transformaciones

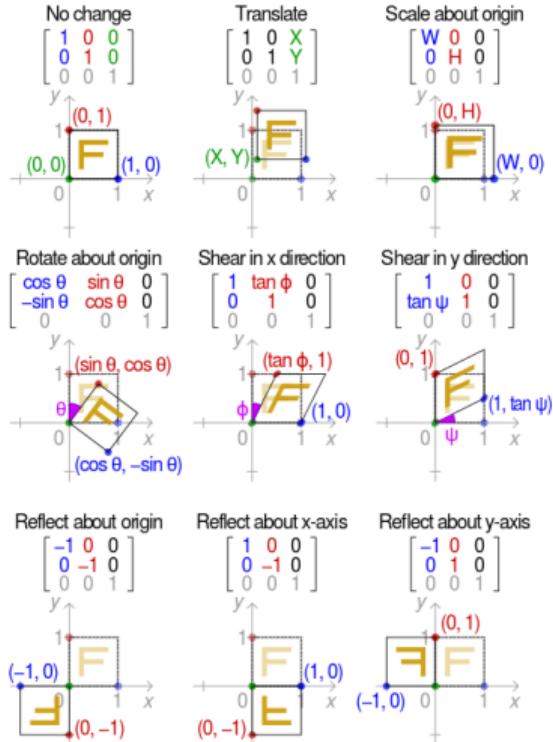


Imagen del usuario de Wikipedia Cmglee (entrada Affine transformation). CC BY-SA 3.0

Acrecentamiento de imágenes: ejemplos

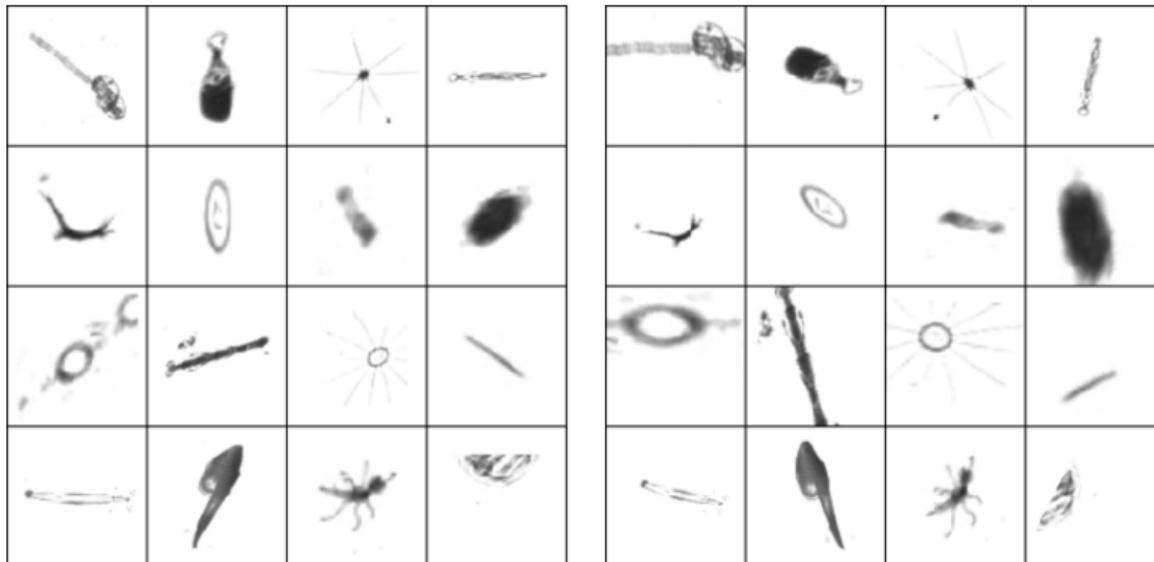


Imagen tomada de Dieleman. Classifying plankton with deep neural networks, 2015.

Acrecentamiento de audio

- Agregar ruido.
- Cambiar la velocidad o tono.
- Enmascaramiento del espectograma¹.

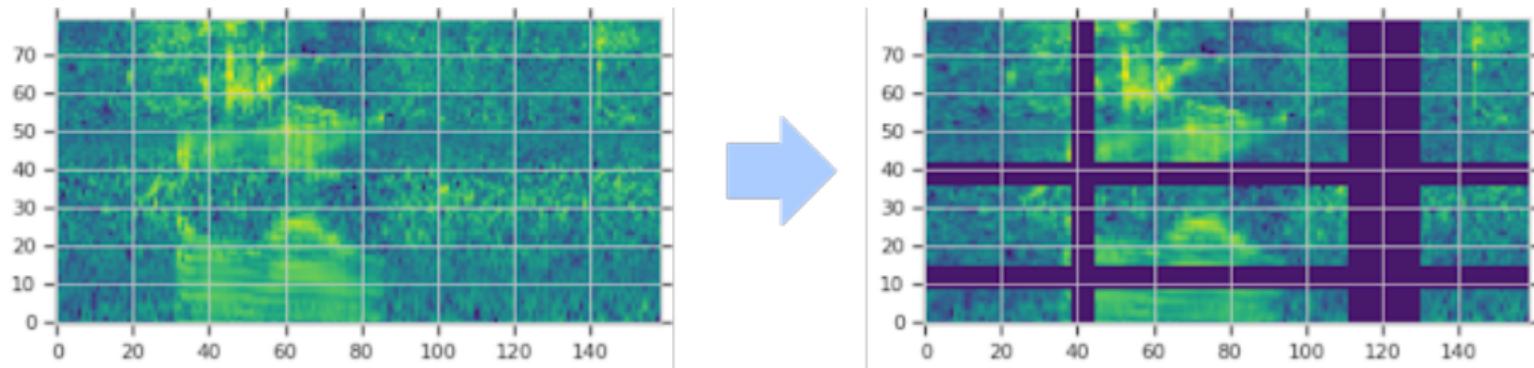


Imagen tomada de <https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html>

¹Daniel S et al. *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition* 2019.

Acrecentamiento en texto

- Símbolos.
 - Insertar/cambiar/quitar símbolos aleatoriamente.
 - Simular errores de teclado.
- Palabras.
 - Cambiar/quitar palabras aleatoriamente o con algún modelo de lenguaje o bolsa de palabras.
 - Cambiar palabras por sinónimos².
 - Cambiar palabras de acuerdo a errores de escritura.

²Zhang et al. *Character-level Convolutional Networks for TextClassification* 2016.

Acrecentamiento en texto: traducción inversa

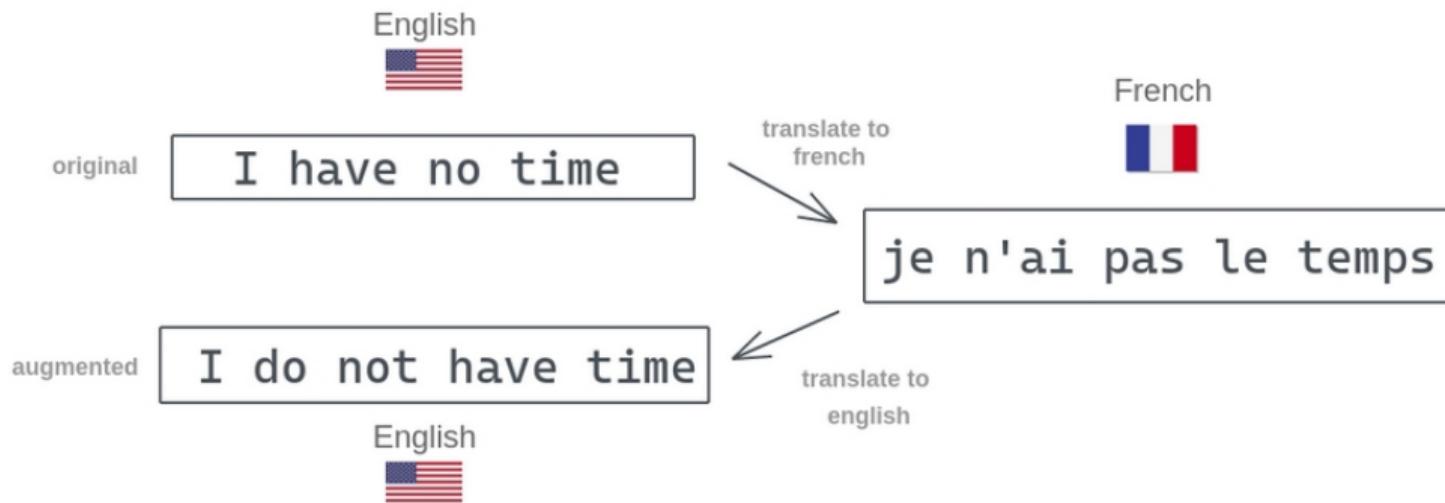
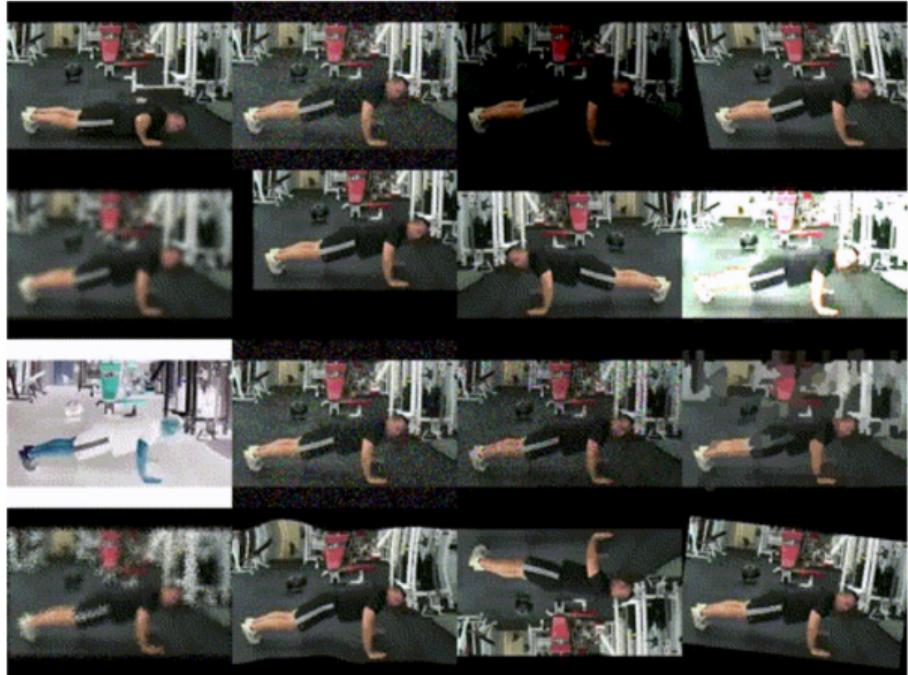


Figure: Back Translation

Imagen tomada de <https://amitness.com/2020/02/back-translation-in-google-sheets>

Acrecentamiento en video

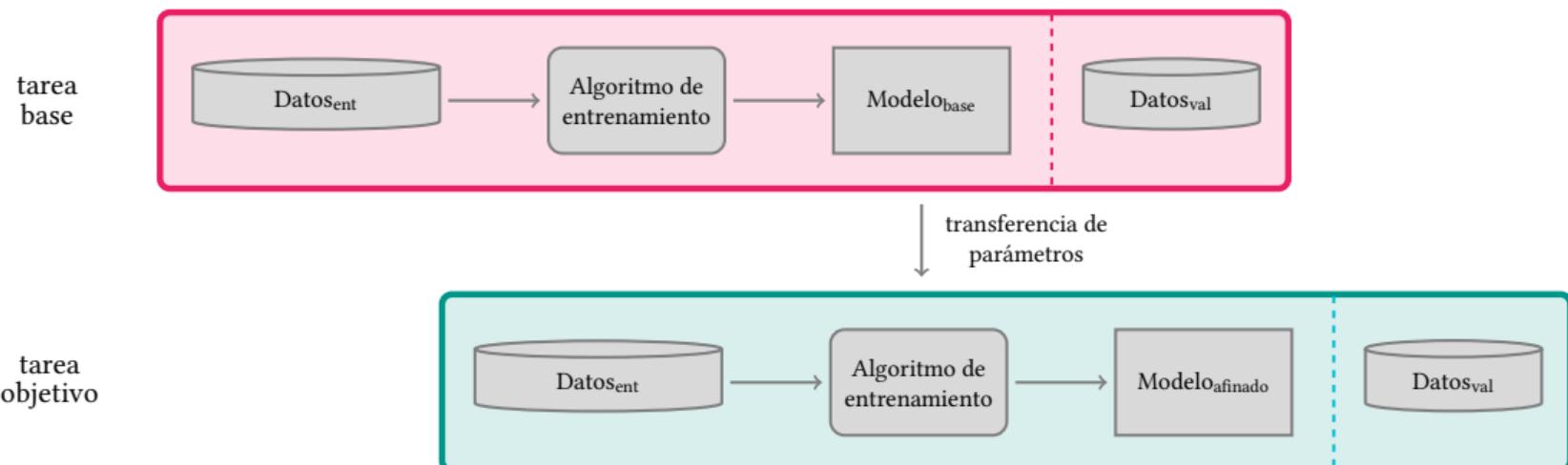
Original Video



Transferencia de conocimiento

Esquema general

- Aprovecha el conocimiento de una tarea base en una tarea objetivo.



Desempeño en distintas tareas

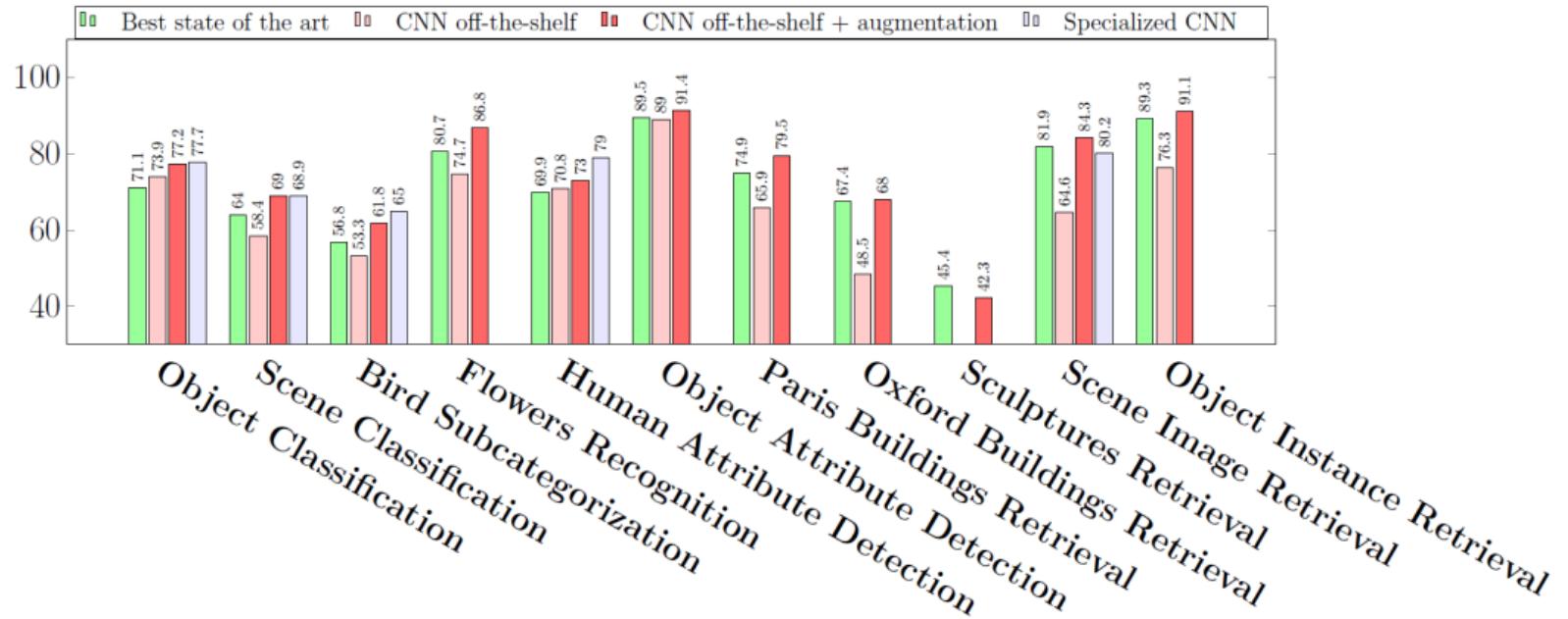


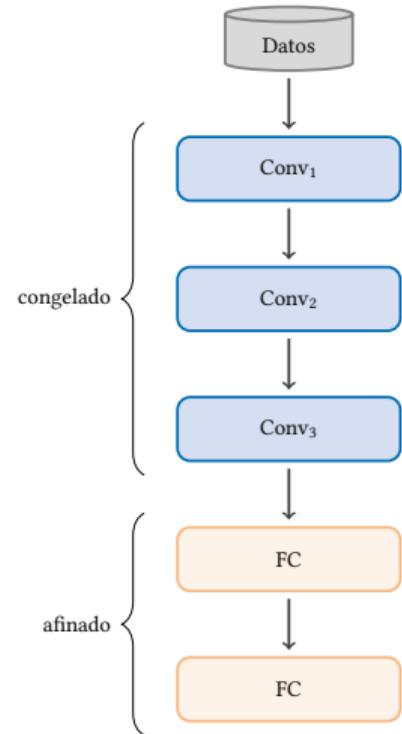
Imagen tomada de Razavian et al. CNN Features off-the-shelf: an Astounding Baseline for Recognition, 2014

Razavian et al. *CNN Features off-the-shelf: an Astounding Baseline for Recognition*. 2014.

Yosinski et al. *How transferable are features in deep neural networks?* 2014.

Transferencia de capas

- Las n primeras capas se pueden congelar o afinar.
 - Congelado: sin actualizaciones durante la retropropagación.
 - Afinado: actualizaciones durante la retropropagación.

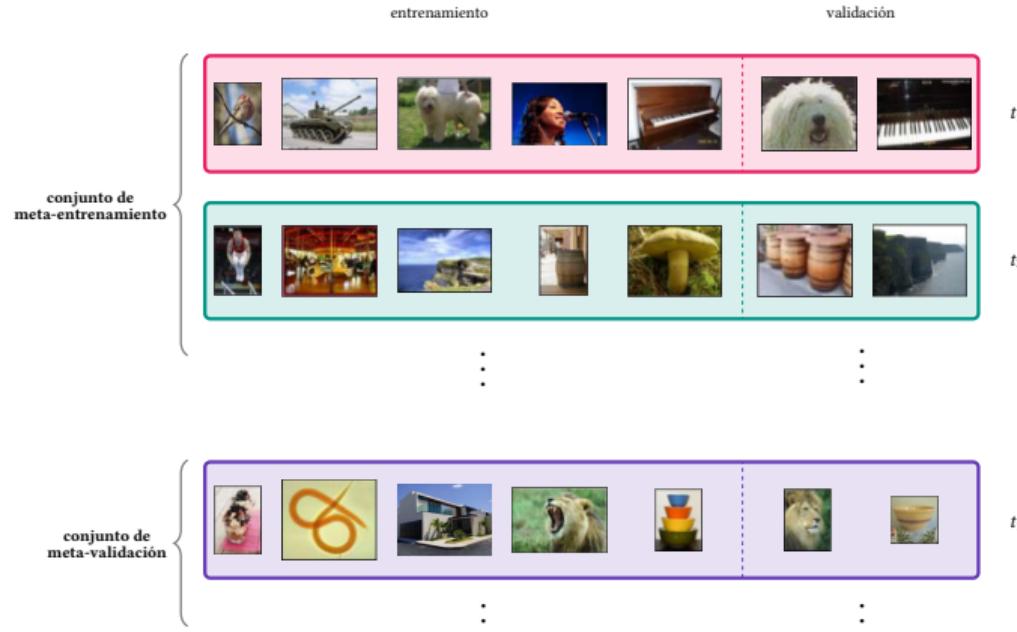


- Estudiar sobre la similitud de las tareas base y objetivo.
 - Diversidad de clases y ejemplos.
 - Desbalanceo.
- Nivel de entrenamiento del modelo preentrenado.
- Procesamiento de datos.
 - Lectura de datos: formatos, canales, bibliotecas, redimensionado.
 - Revisar con cuidado la normalización usada por el modelo original.
- Estrategia de entrenamiento.
 - Solo la última capa.
 - Toda la red.
 - Aprendizaje diferencial ³.

³Singh et al. *Layer-Specific Adaptive Learning Rates for Deep Networks*. 2018.

Meta aprendizaje

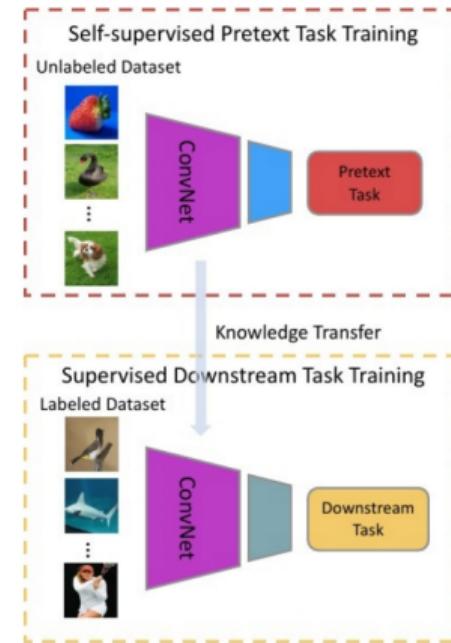
- Familia de técnicas enfocadas adaptarse rápidamente a nueva información.



Aprendizaje autosupervisado

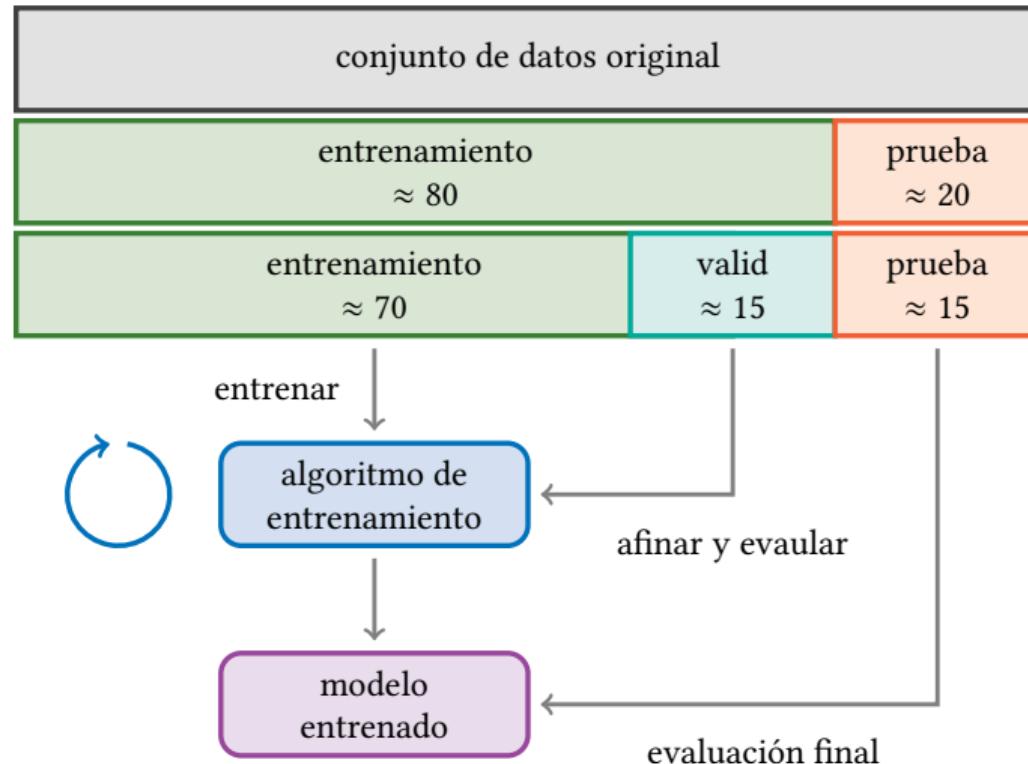
- Aprender de forma supervisada en datos sin etiquetas.

- Tareas de naturaleza autosupervisada
- Preentrenamiento y transferencia de conocimiento



Partición de los datos

Partición de los datos



Sobreajuste

Como identificar el sobreajuste

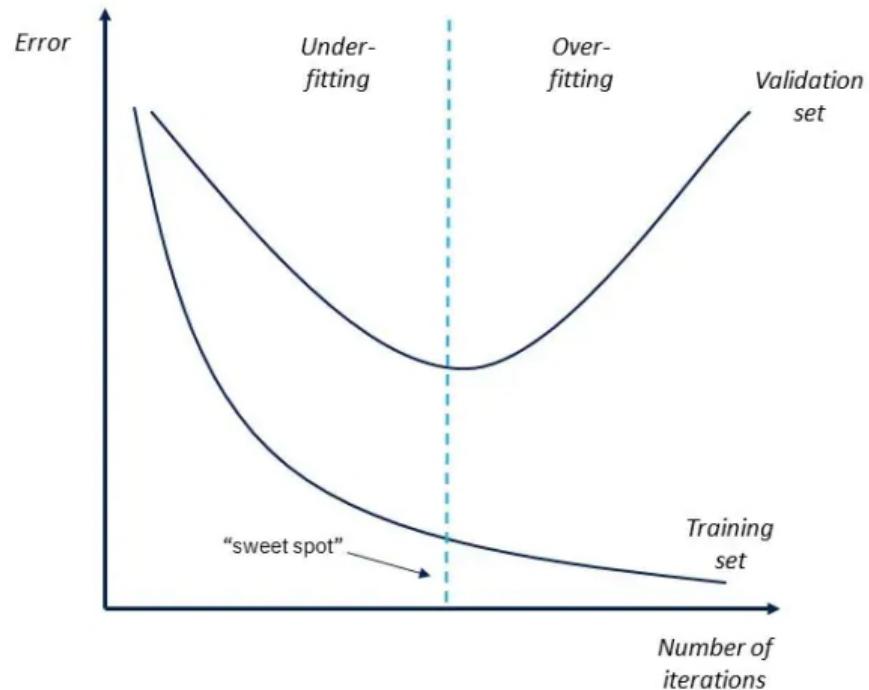


Imagen tomada de <https://www.ibm.com/cloud/learn/overfitting>

Estrategias para mitigar el sobreajuste

- Estrategias para reducir error de generalización:
 - Penalización de función de error (o función de pérdida).
 - Adición de ruido a entradas, salidas y/o parámetros.
 - Ensambles.
 - Paro temprano.
 - Aprendizaje de múltiples tareas.
 - Dropout.
 - Normalización por lotes.

Penalización de pesos y sesgos con norma ℓ_1 y ℓ_2

- Norma ℓ_1

$$J(\boldsymbol{\theta}) = - \sum_{i=1}^n \{y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\} + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_1$$

- Norma ℓ_2

$$J(\boldsymbol{\theta}) = - \sum_{i=1}^n \{y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\} + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

Paro temprano

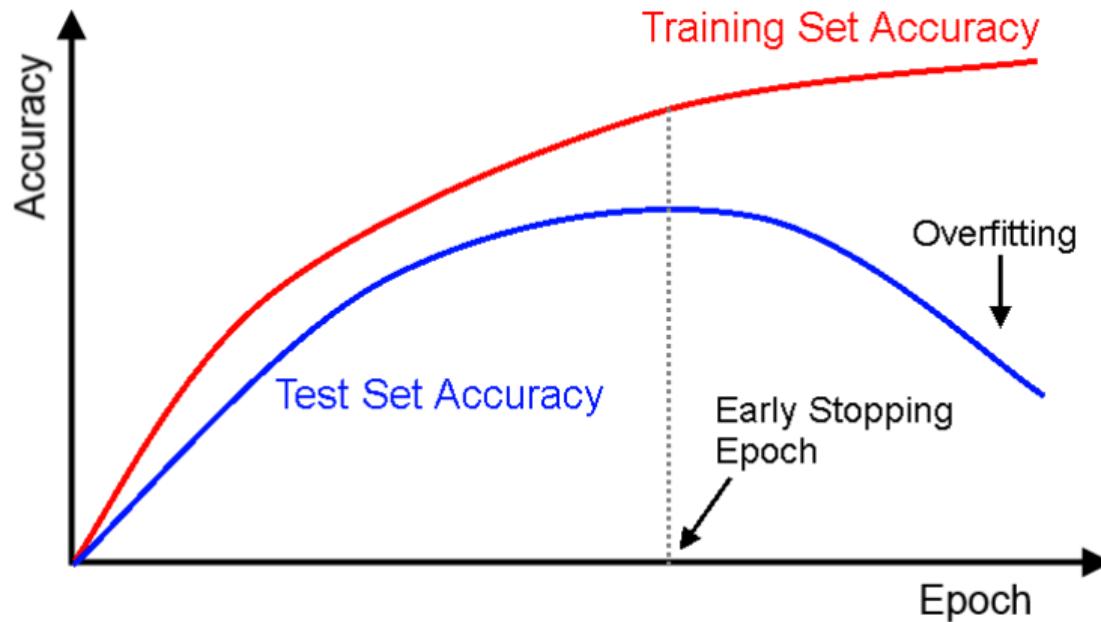
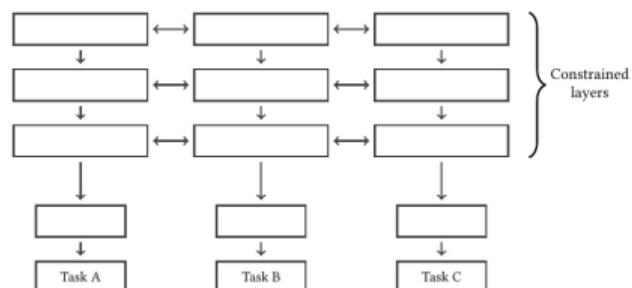


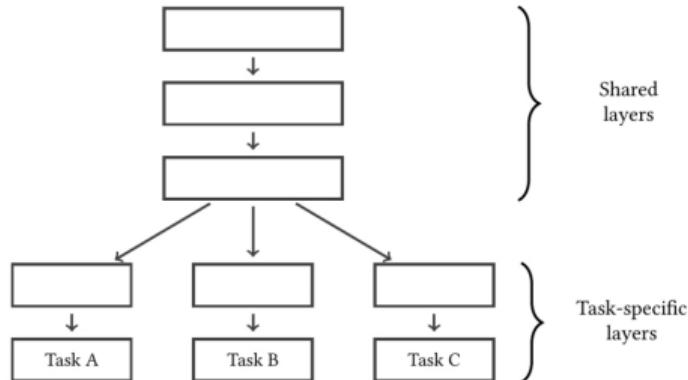
Imagen tomada de <https://deeplearning4j.org/earlystopping>

Aprendizaje de multitarea

- Aprendizaje simultaneo de dos o más tareas.

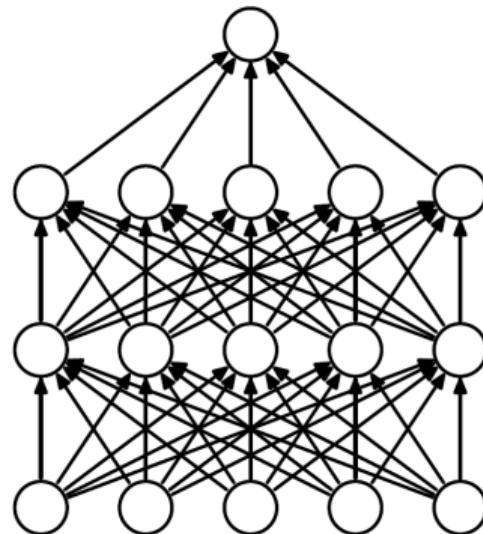


Compartición suave
de parámetros

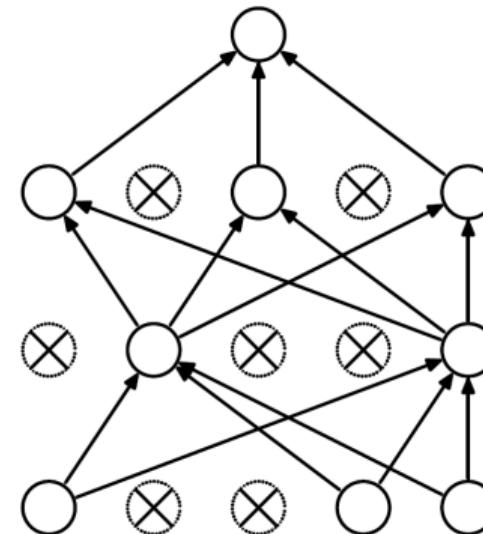


Compartición dura
de parámetros

- Se desactivan neuronas de forma aleatoria⁴ para evitar co-adaptación



(a) Standard Neural Net



(b) After applying dropout.

Imagen tomada de Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting, 2014

⁴La probabilidad es típicamente de 0.5.

Deserción como ensamble

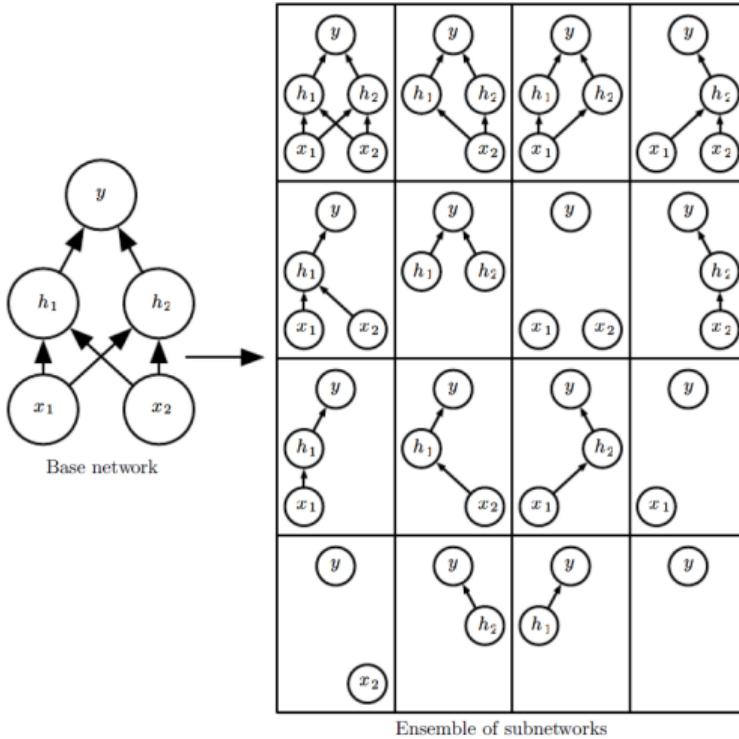


Imagen tomada de Goodfellow et al. Deep Learning, 2016

Mejorando convergencia

Normalización por lote

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots m\}$;

Parameters to be learned: γ, β

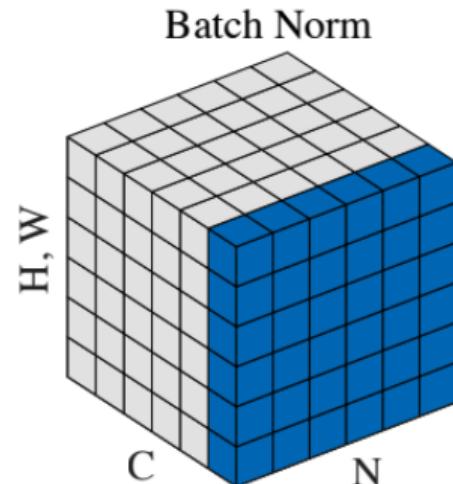
Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$



Beneficios de la normalización por lotes

- Acelera el entrenamiento.
- Permite tasas de aprendizaje más grandes.
- Facilita la inicialización de pesos.
- Hace posible usar funciones de activación saturadas (por ej. sigmoide).
- Actúa como un tipo de regularizador.
- Facilita la creación de redes profundas.

Inicialización de pesos (I)

- Números aleatorios de distribución gaussiana con media 0 y varianza 0.01.
 - Funciona en redes pequeñas.
 - Para redes profundas activaciones tienden a volverse 0.
- Números aleatorios de distribución gaussiana con media 0 y varianza 1.
 - Genera saturación de las neuronas y gradientes se vuelven 0.

Inicialización de pesos (II)

- Para una capa con n_e entradas y n_s salidas
 - Uniforme de Glorot y Bengio (2010).

$$\theta \sim \mathcal{U} \left[-\sqrt{\frac{6}{n_e + n_s}}, \sqrt{\frac{6}{n_e + n_s}} \right]$$

- Normal de Glorot y Bengio (2010).

$$\theta \sim \mathcal{N} \left(0, \frac{2}{n_e + n_s} \right)$$

- Normal de He et al. (2015).

$$\theta \sim \mathcal{N} \left(0, \frac{2}{n_e} \right)$$