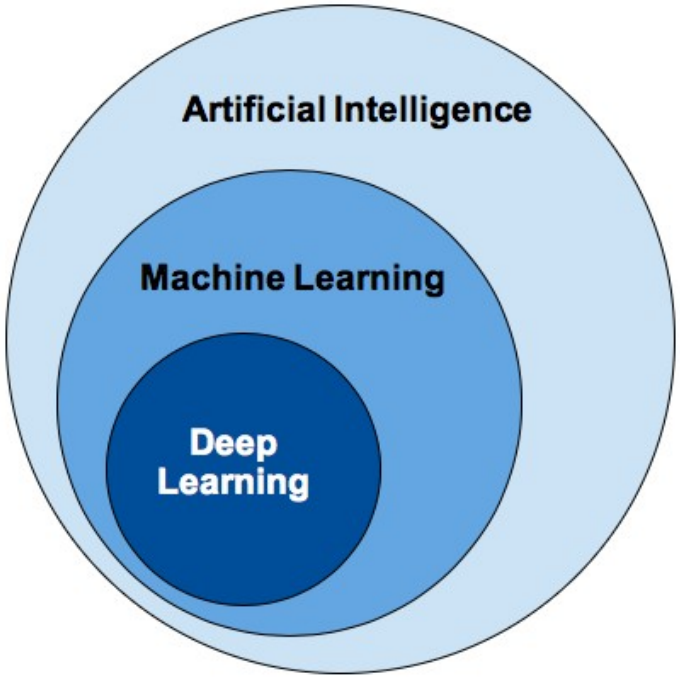
Redes profundas y redes convolucionales



Basado en charlas de Lucas C. Uzal y otros

“Revolution of Depth”

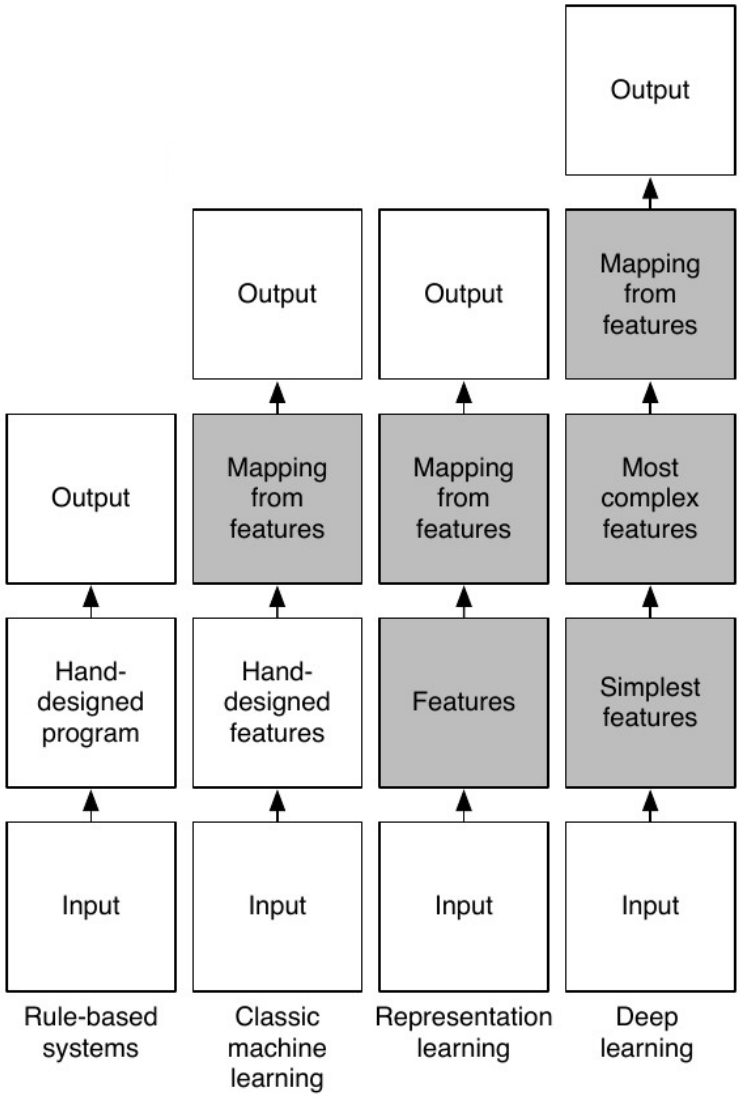
Situación a mitad de los “2000”

- Machine Learning está establecido como campo de investigación con alto potencial tecnológico
- Las SVM dominan el campo
- Tipo de solución basado en el conocimiento de expertos:
  - Medir cosas (features) sobre los datos, sugeridas por conocimiento del problema
  - Aplicar un método de ML a esos features para decidir
- Las redes neuronales están en un “invierno”:
  - Se sabe que las redes neuronales son aproximadores universales.
  - Se sabe que las redes tienen el potencial de aprender su propia representación.
  - Problemas irresolubles: limitaciones para resolver problemas reales, dificultad para utilizarlas, falta de garantías de convergencia, lentitud de los entrenamiento

“Revolution of Depth”

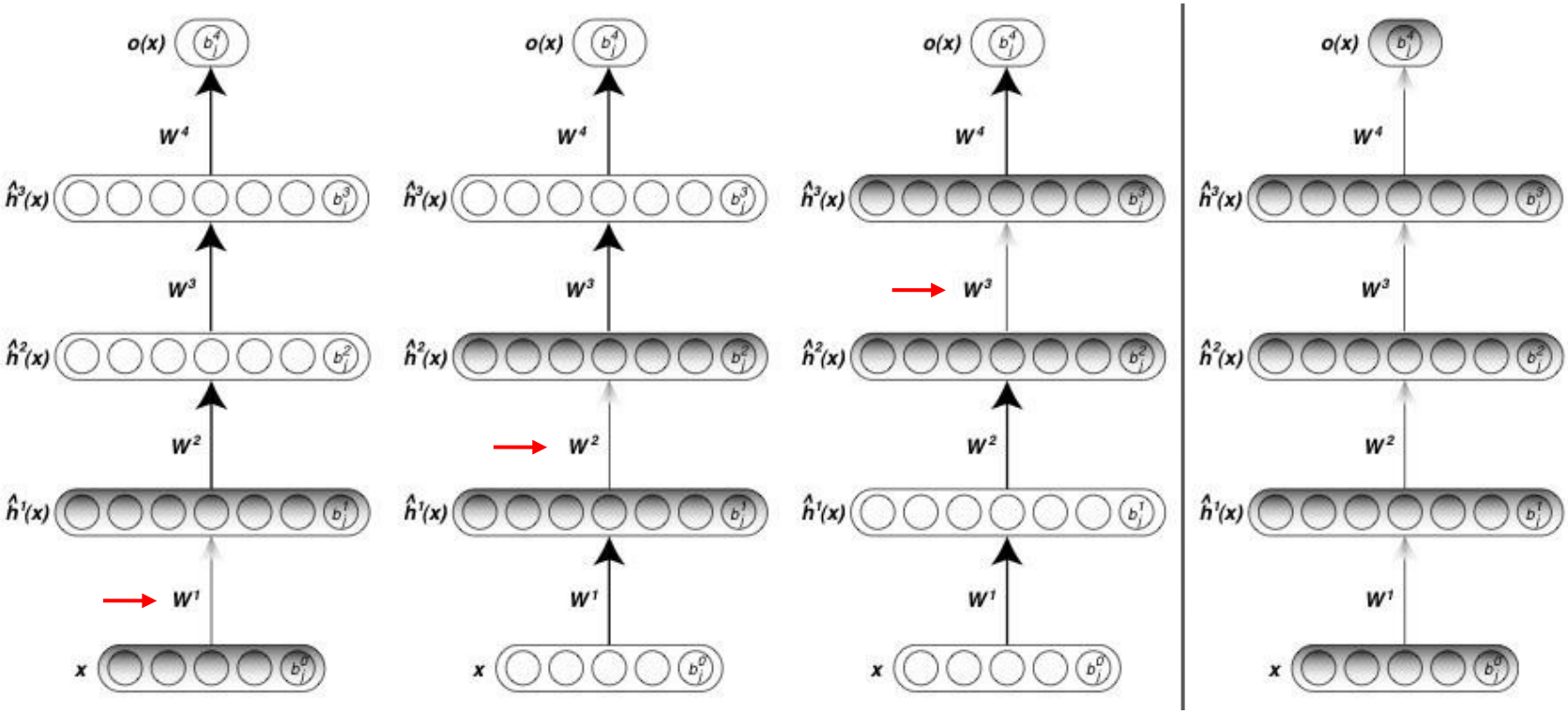
Por qué no seguir el mismo camino?

- “Aprender” las reglas produjo sistemas que solucionan las deficiencias de los sistemas expertos tradicionales.
- “Aprender” los features podría solucionar los problemas actuales.
- Si los métodos actuales no aprenden bien, por qué no usar métodos que puedan ir de lo simple a lo complejo?



No se sabía cómo hacer que las redes aprendan de esta forma!

2006: Greedy layer-wise training of deep networks



Hinton, Geoffrey E, and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507.

Hinton, Geoffrey, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation* 18.7 (2006): 1527-1554.

Bengio, Yoshua et al. "Greedy layer-wise training of deep networks." *Advances in neural information processing systems* 19 (2007): 153.

“Revolution of Depth”

Primera solución al problema de entrenar una red profunda.

- Hinton muestra que se puede encontrar una primera solución entrenando de manera no supervizada, y afinando la solución (fine tunning) después, pero en redes muy particulares.
- Resurgen las redes como método, y hay un avance exponencial que lleva a la actual revolución tecnológica.
- En pocos años se encuentra como entrenar redes generales, y como extender todo lo que ya se sabía hacer a modelos profundos.

3 claves: Datos (internet), Potencia (GPUs), Métodos (activaciones, regularización).

Google AI algorithm masters ancient game of Go

Deep-learning software defeats human professional for first time. x

January 2016

A computer has beaten a human professional for the first time at Go — an ancient board game that has long been viewed as one of the greatest challenges for artificial intelligence (AI)

"We pass in the board position as a 19 × 19 image and use convolutional layers to construct a representation of the position."

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

nature International weekly journal of science

NATURE | LETTER

日本語要約

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Affiliations | Contributions | Corresponding authors

Nature **542**, 115–118 (02 February 2017) | doi:10.1038/nature21056

Received 28 June 2016 | Accepted 14 December 2016 | Published online 25 January 2017

LA NACION

Un nuevo algoritmo es capaz de diagnosticar cáncer de piel con la misma eficacia que un dermatólogo

Hicieron que un sistema de inteligencia artificial analizara miles de imágenes para aprender a reconocer melanomas; por ahora es un prototipo

n/Flickr bajo licencia CC 2.0

MIÉRCOLES 25 DE ENERO DE 2017 • 18:04

Un nuevo algoritmo desarrollado por la universidad de Stanford es capaz de diagnosticar un cáncer de piel con la misma precisión que un médico humano. No es el

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.


A.M.
TURING
AWARD
2018
acm
YOSHUA BENGIO,
GEOFFREY E. HINTON
AND YANN LECUN
For conceptual and engineering
breakthroughs that have made
deep neural networks a critical
component of computing

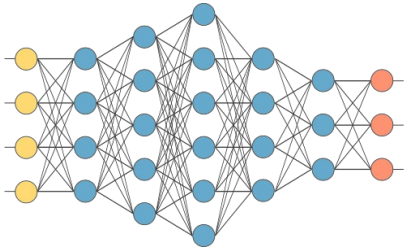
Grandes áreas

Deep Learning

Datos Secuenciales

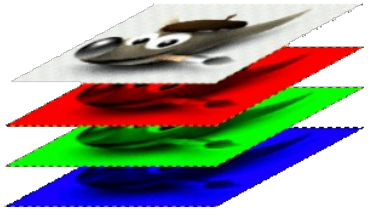
Imágenes

REDES NEURONALES



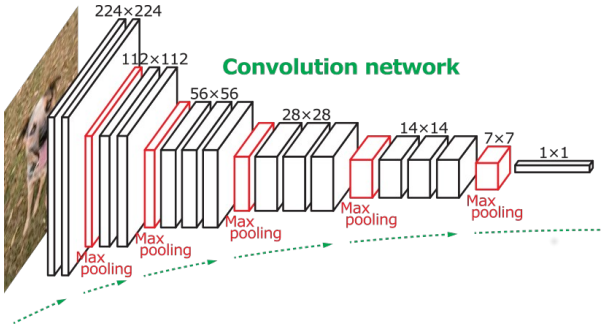
+

IMÁGENES



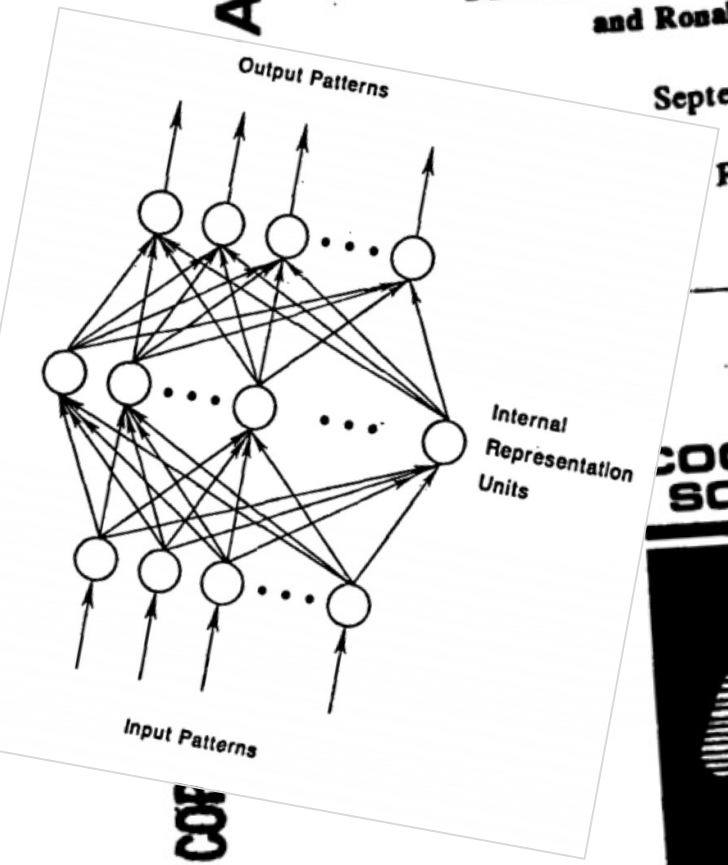
=

REDES NEURONALES CONVOLUCIONALES



Algo de historia

Primeras redes neuronales



A164 453

LEARNING INTERNAL REPRESENTATIONS BY ERROR PROPAGATION

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams

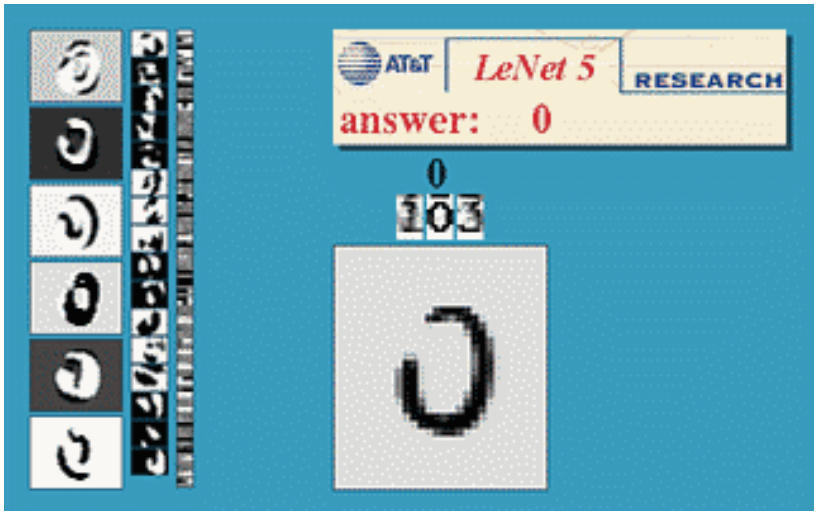
September 1985

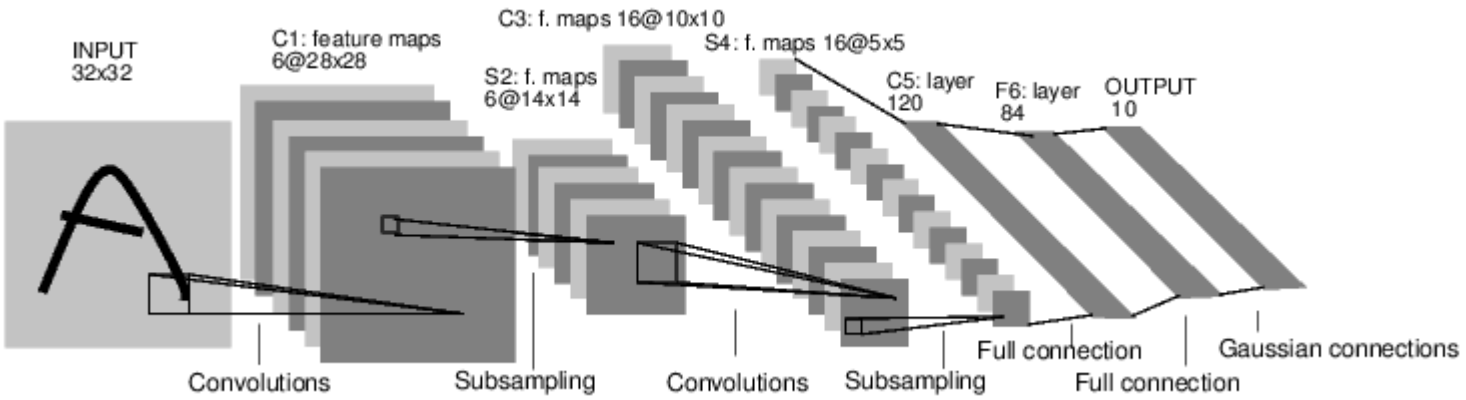
Report 8506

DTIC ELECTE FEB 2 0 1986 D

Geoffrey E. Hinton

Primera red convolucional

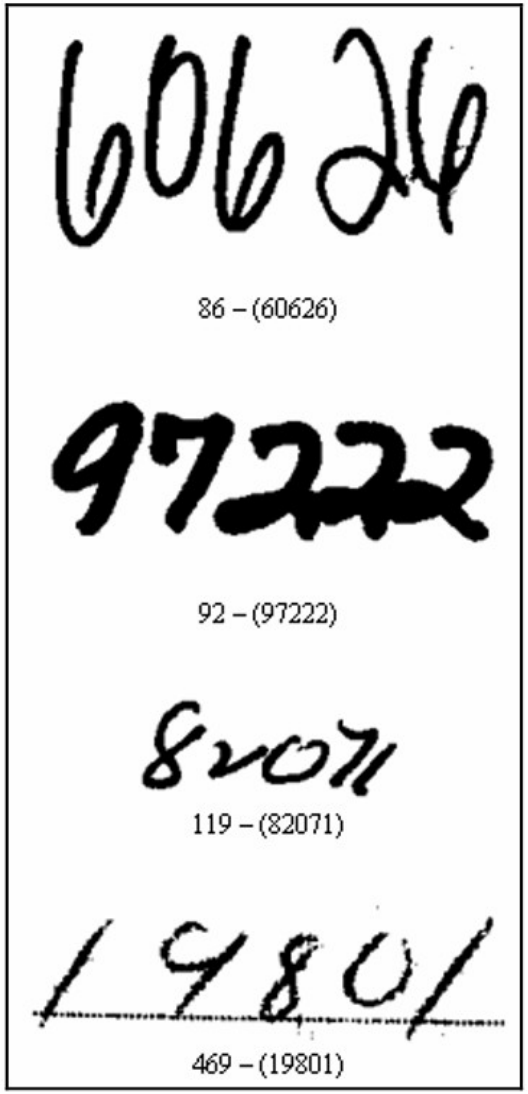


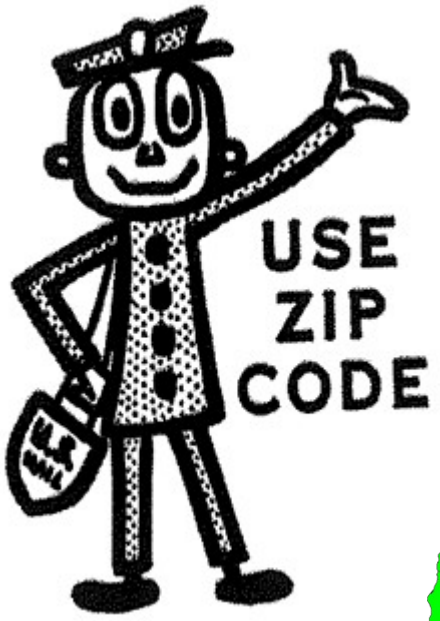


Yann LeCun

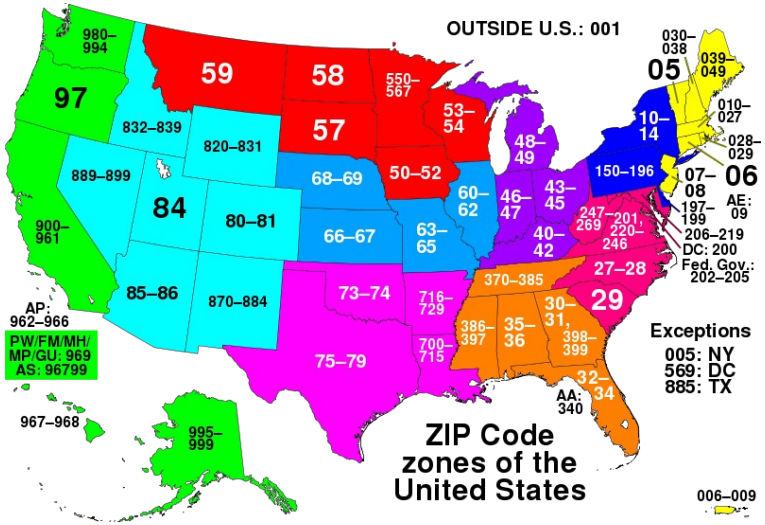
LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Lectura de código postal



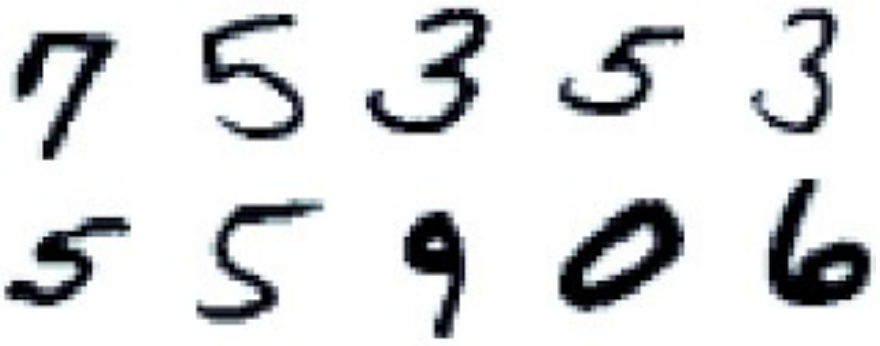


the90s



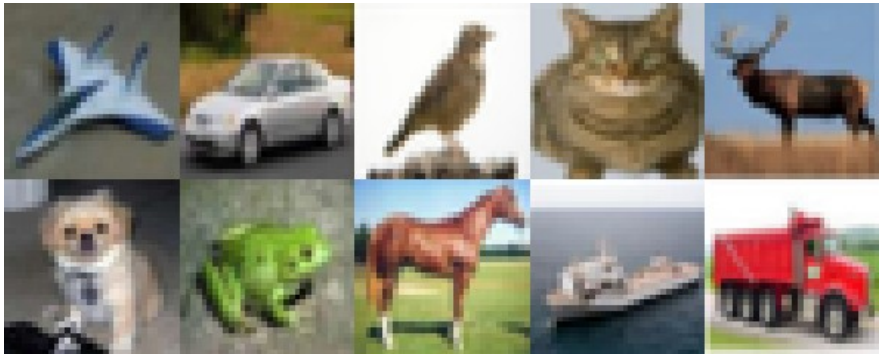
El problema de la resolución

28x28



MNIST

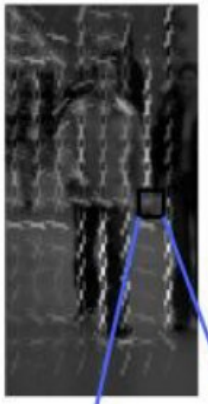
32x32



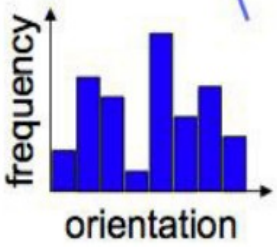
CIFAR-10

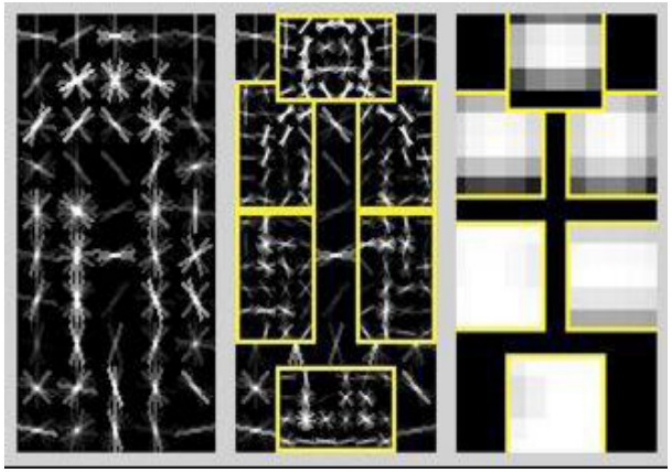
96x96

STL-10



Histogram of Gradients (HoG) Dalal & Triggs, 2005





Deformable Part Model Felzenswalb, McAllester, Ramanan, 2009

Dense grid descriptor: HOG, LBP

Coding: local coordinate, super-vector

Pooling, SPM

Linear SVM

2010

IMAGENET Large Scale Visual Recognition Challenge

Steel drum

The Image Classification Challenge:

1,000 object classes

1,431,167 images

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

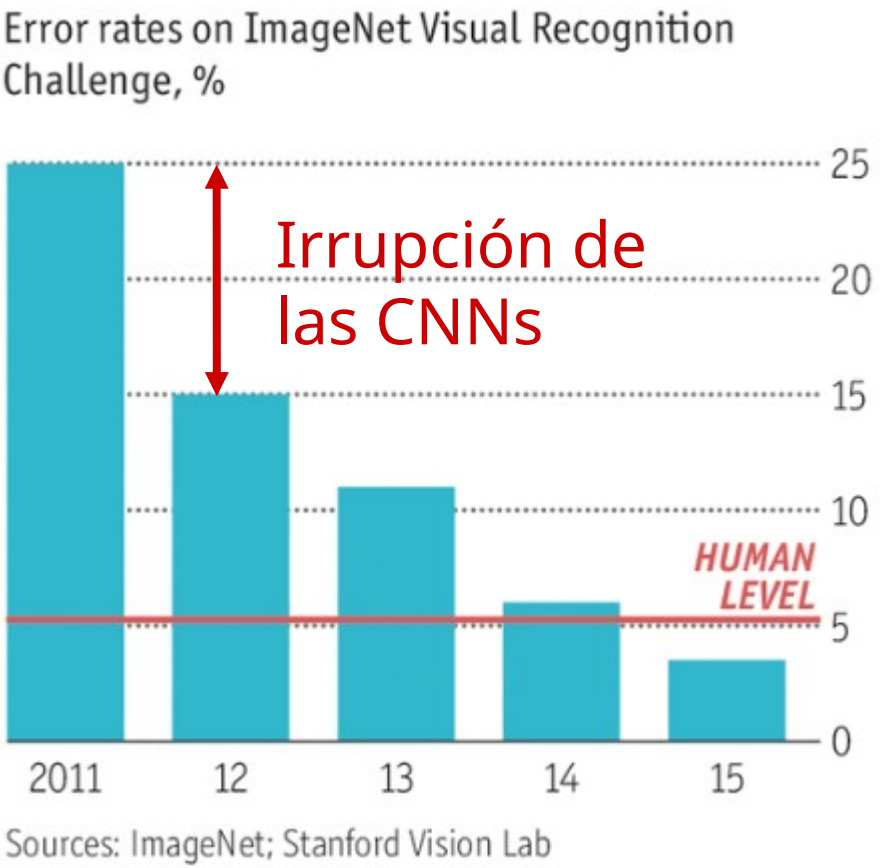
✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge.International Journal of Computer Vision, 115(3), 211-252.

Las redes convolucionales ganan la carrera



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Yann LeCun

Geoffrey E. Hinton

Las redes convolucionales ganan la carrera

Year 2010

NEC-UIUC

Dense grid descriptor: HOG, LBP

Coding: local coordinate, super-vector

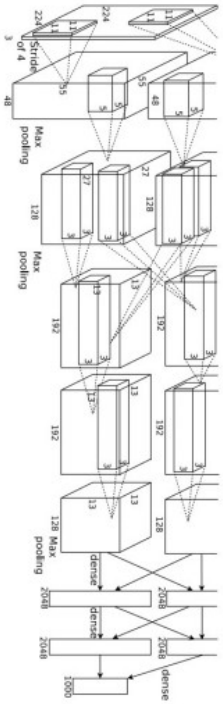
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet

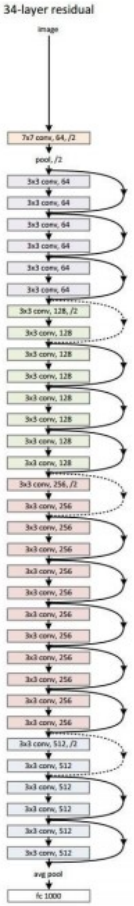
[Szegedy arxiv 2014]

VGG

[Simonyan arxiv 2014]

Year 2015

MSRA



“Revolution of Depth”

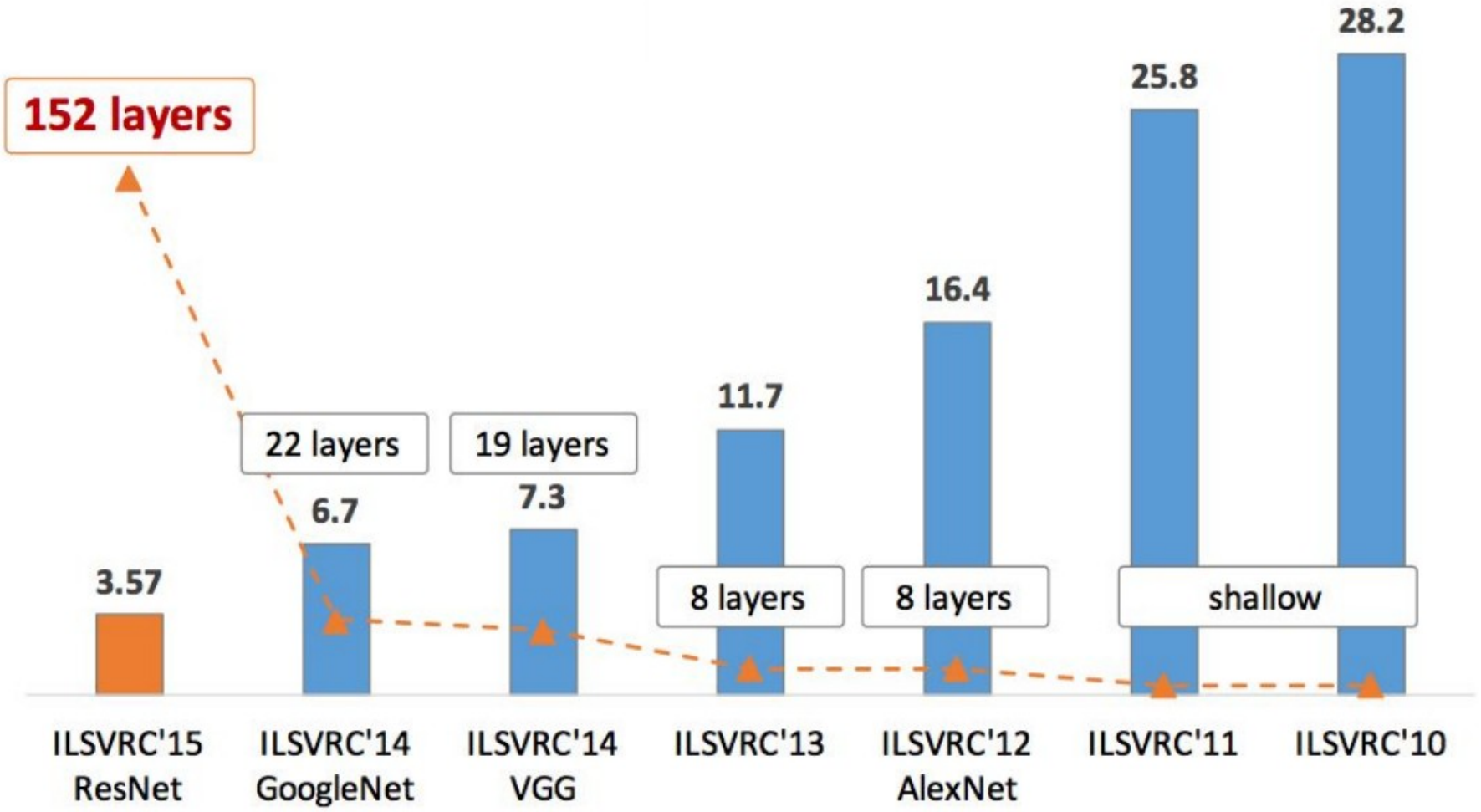
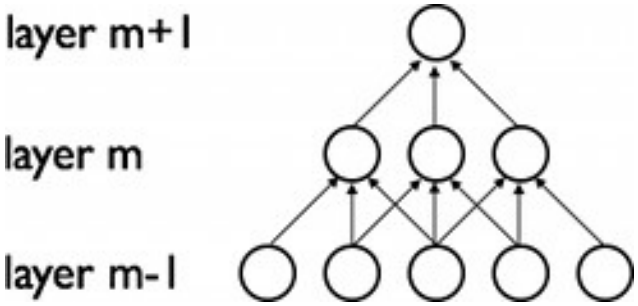


Figure copyright Kaiming He, 2016.

Redes convolucionales

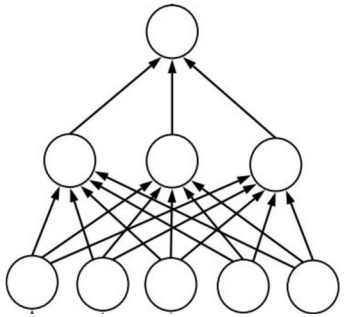
Convolutional Neural Networks

Sparse connectivity

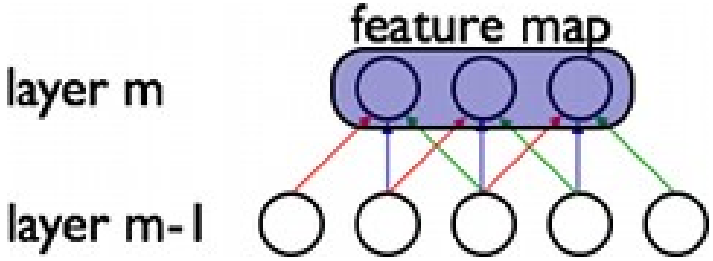


versus

Dense connectivity



Shared weights

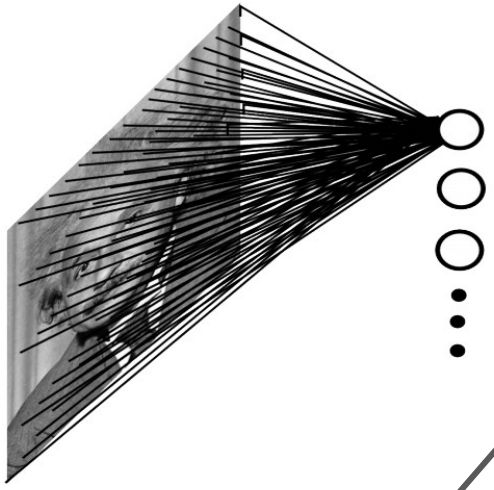


http://deeplearning.net/tutorial/lenet.html

Convolutional Neural Networks

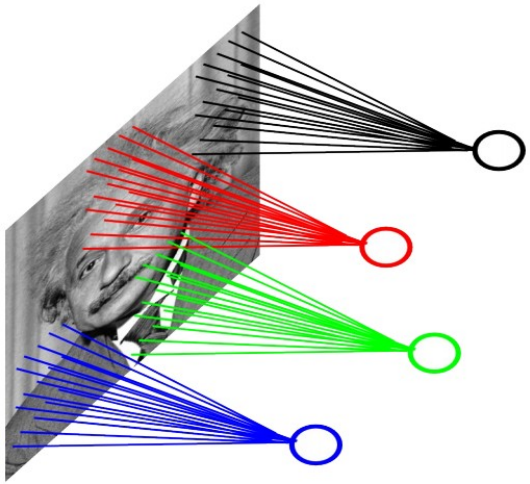
dot product + bias

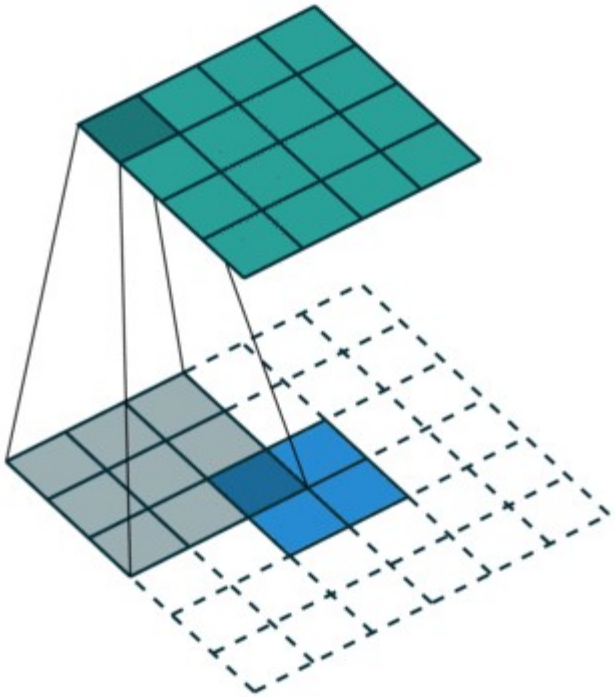
$$h_k = \mathrm{ReLU}\left(W_k \cdot x + b_k\right)$$



convolution + bias

$$h^k_{ij} = \mathrm{ReLU}\left((W_k * x)_{ij} + b_k\right)$$





http://www.cs.nyu.edu/~yann/talks/lecun-ranzato-icml2013.pdf

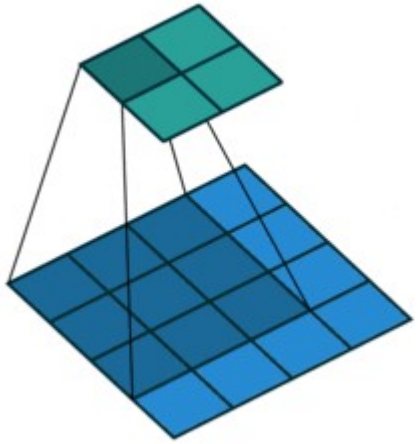
Convolutional Neural Networks

Por qué?

- **Invariancia traslacional.** La visión responde igual en cualquier lugar, una cara es una cara en cualquier lugar de la imagen
- **Localidad.** Entender qué hay en una imagen depende del lugar en que se concentra la atención.
- **Reducir sobreajuste.** Al tener menos pesos ajustables, se reduce la posibilidad de aprender particularidades.

Convolutional Neural Networks

Ejemplo



Input

0	1	2
3	4	5
6	7	8

*

Kernel

0	1
2	3

=

Output

19	25
37	43

Padding y Stride

Input

0	0	0	0	0
0	0	1	2	0
0	3	4	5	0
0	6	7	8	0
0	0	0	0	0

*

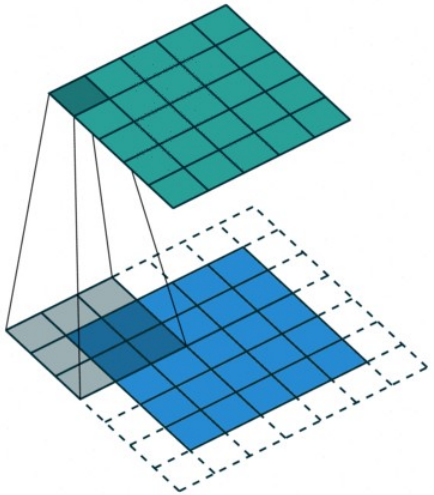
Kernel

0	1
2	3

=

Output

0	3	8	4
9	19	25	10
21	37	43	16
6	7	8	0



Stride is the number of "unit" the kernel shifted per slide over rows/columns.

E.g., Strides of 3 for height and 2 for width

Input

0	0	0	0	0
0	0	1	2	0
0	3	4	5	0
0	6	7	8	0
0	0	0	0	0

*

Kernel

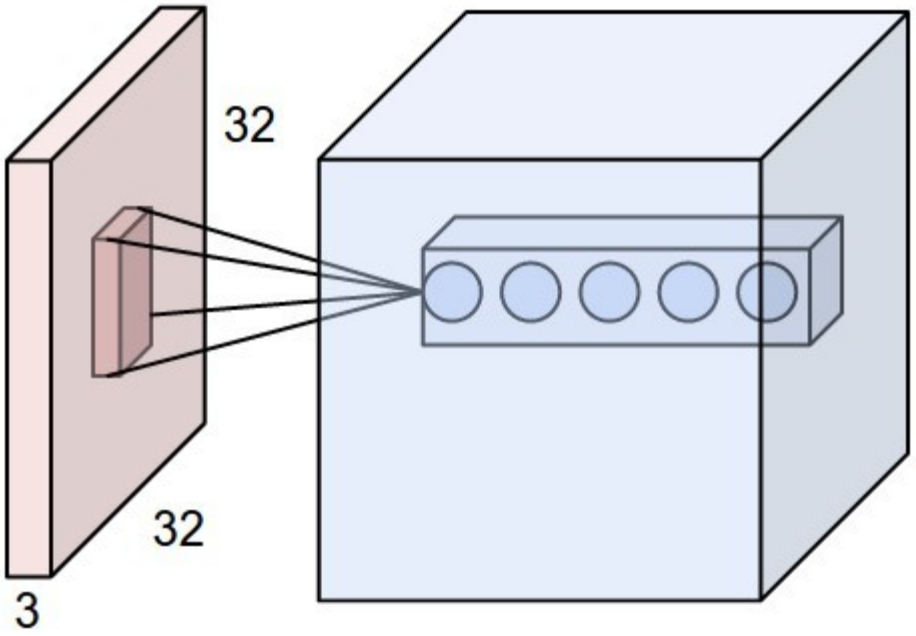
0	1
2	3

=

Output

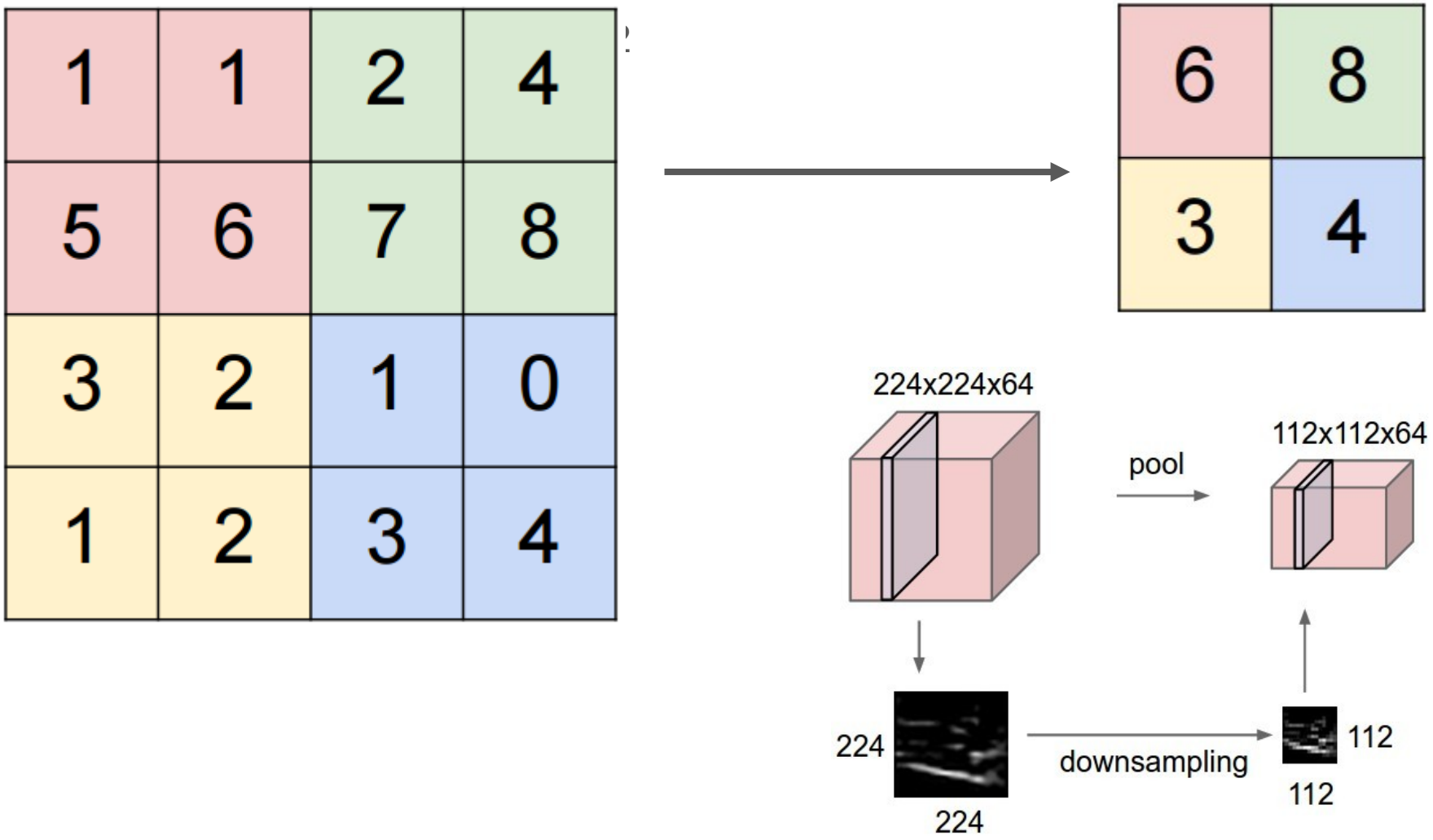
0	8
6	8

Una capa convolucional transforma el volumen

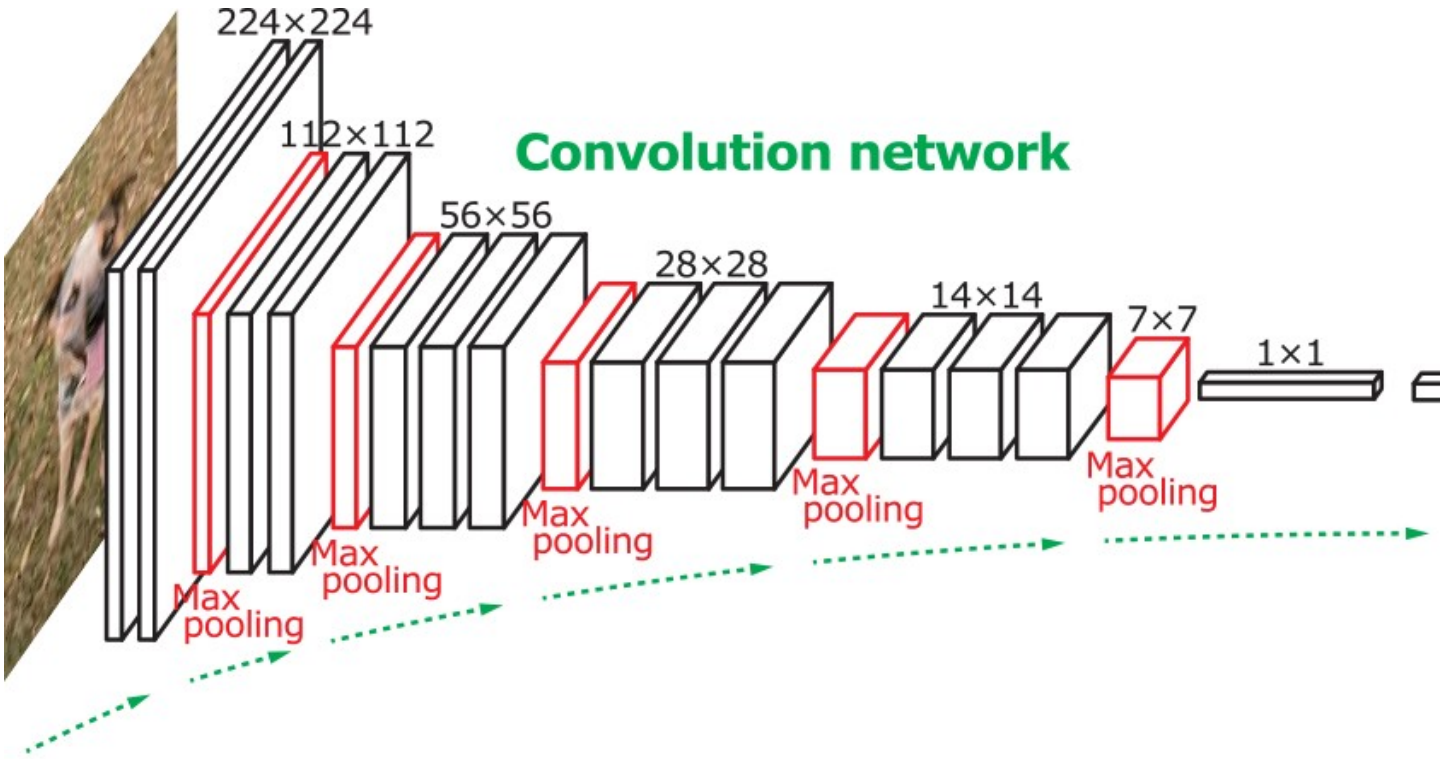


Entrada: imagen, 32x32, 3 canales RGB: 32x32x3
Filtro: tamaño 3x3, x 3 canales.
Salida: aplico 5 filtros convolucionales distintos,
obtengo por cada uno una “imagen” de 32x32: 32x32x5

Reducción de volumen: Pooling



Ejemplo: VGG net ("versión D")



Input 224x224
conv3-64
conv3-64
maxpool
conv3-128
conv3-128
maxpool
conv3-256
conv3-256
conv3-256
maxpool
conv3-512
conv3-512
conv3-512
maxpool
conv3-512
conv3-512
conv3-512
maxpool
FC-4096
FC-4096
FC-1000
softmax

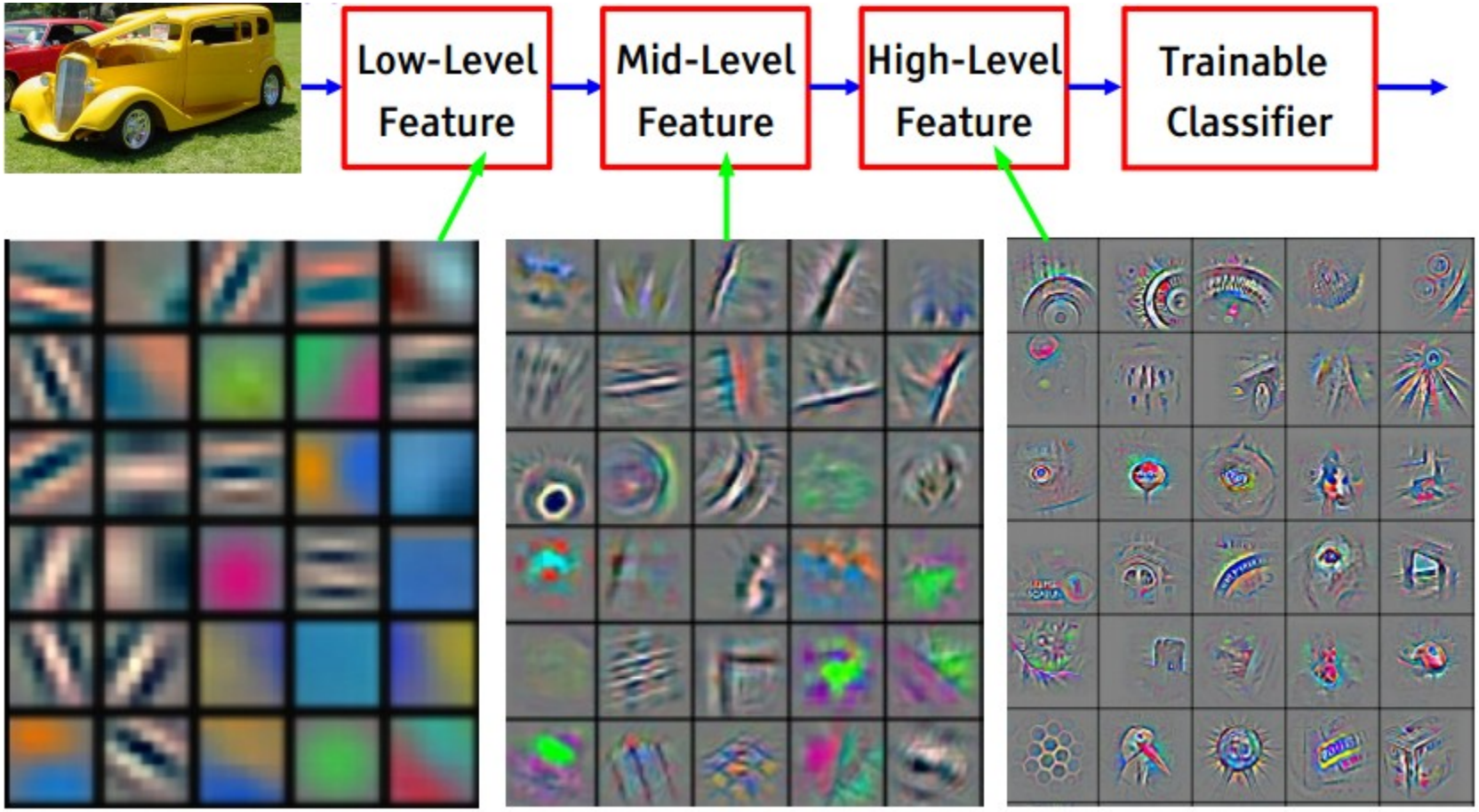
dataset: ImageNet (ILSVRC 2013)

¿Cómo se entrena una CNN?

Backpropagation.

- **Derivación automática.**
- **Learning rate adaptativos.**
- **Minibatch.**
- **Unidades Relu para mantener el gradiente.**

¿Qué ve una CNN ya entrenada?

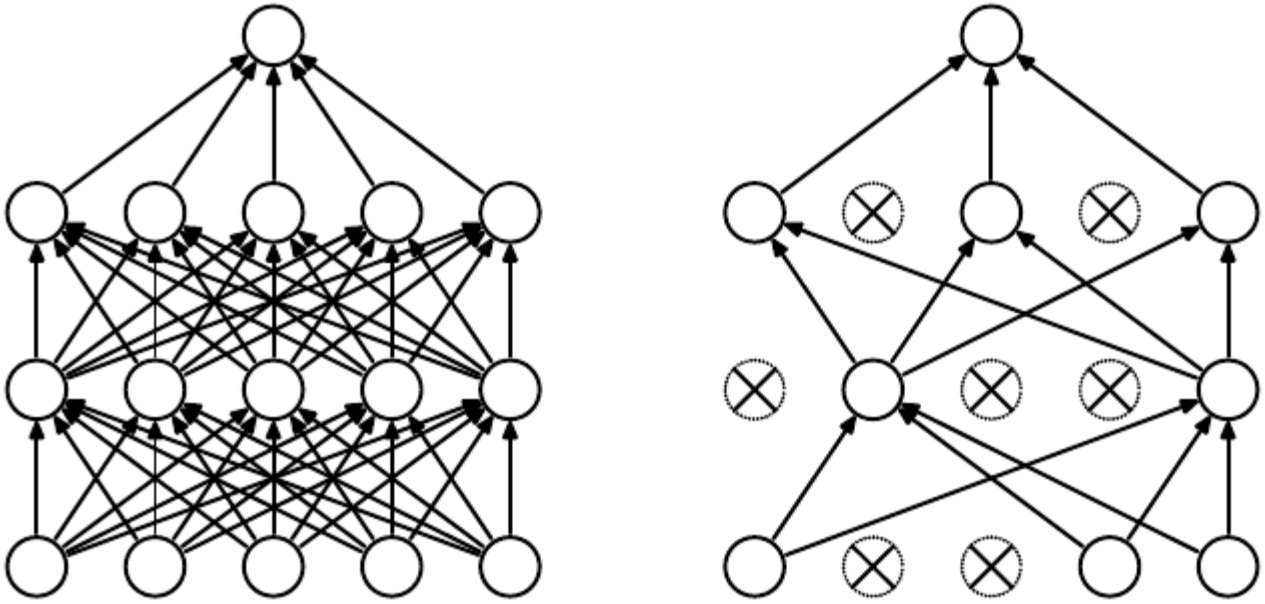


Zeiler, Matthew D, and Rob Fergus. "Visualizing and understanding convolutional neural networks." arXiv preprint arXiv:1311.2901 (2013).

Regularización

Dropout y Data augmentation

Dropout Neural Net Model

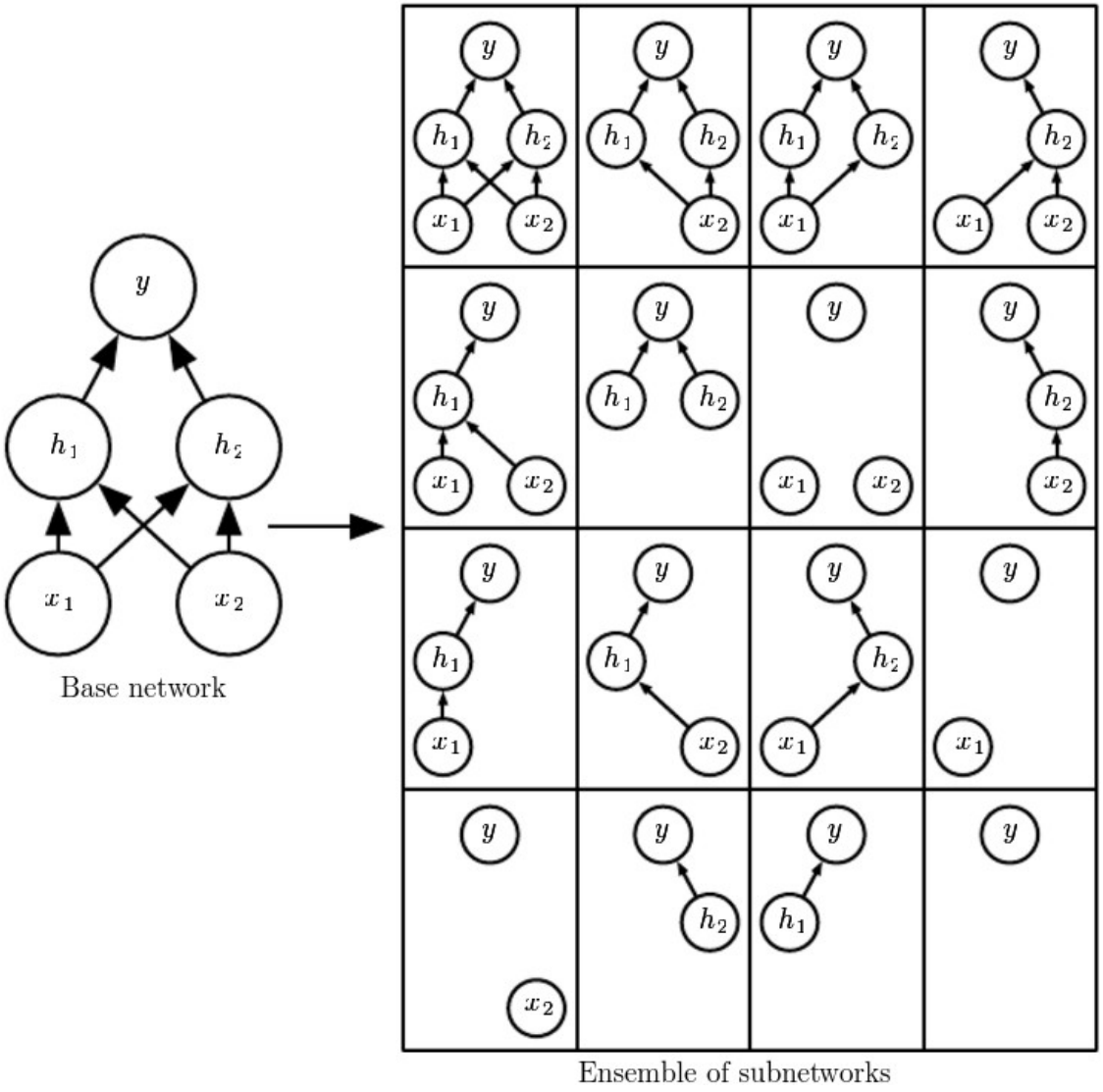


(a) Standard Neural Net (b) After applying dropout.

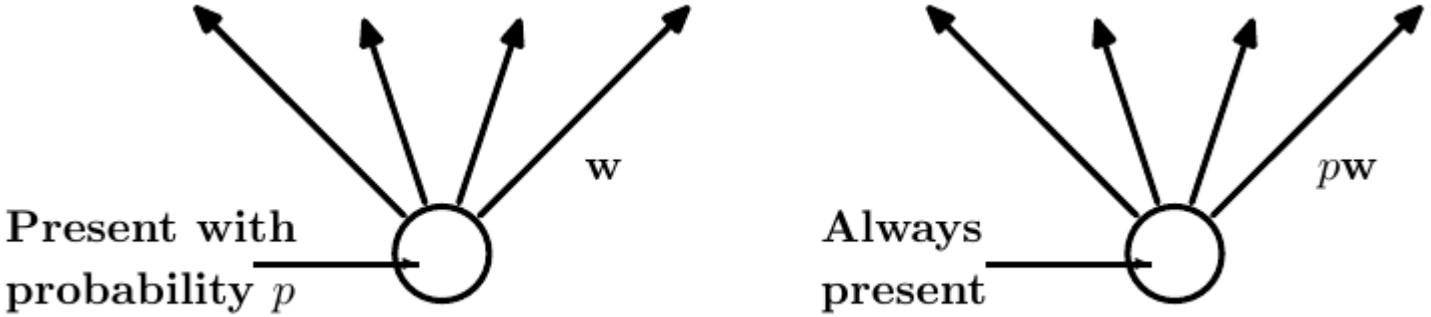
Dropout Neural Net Model. **Left**: A standard neural net with 2 hidden layers. **Right**: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

Srivastava, Nitish, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (2014): 1929-1958.

Dropout Neural Net Model



Dropout units

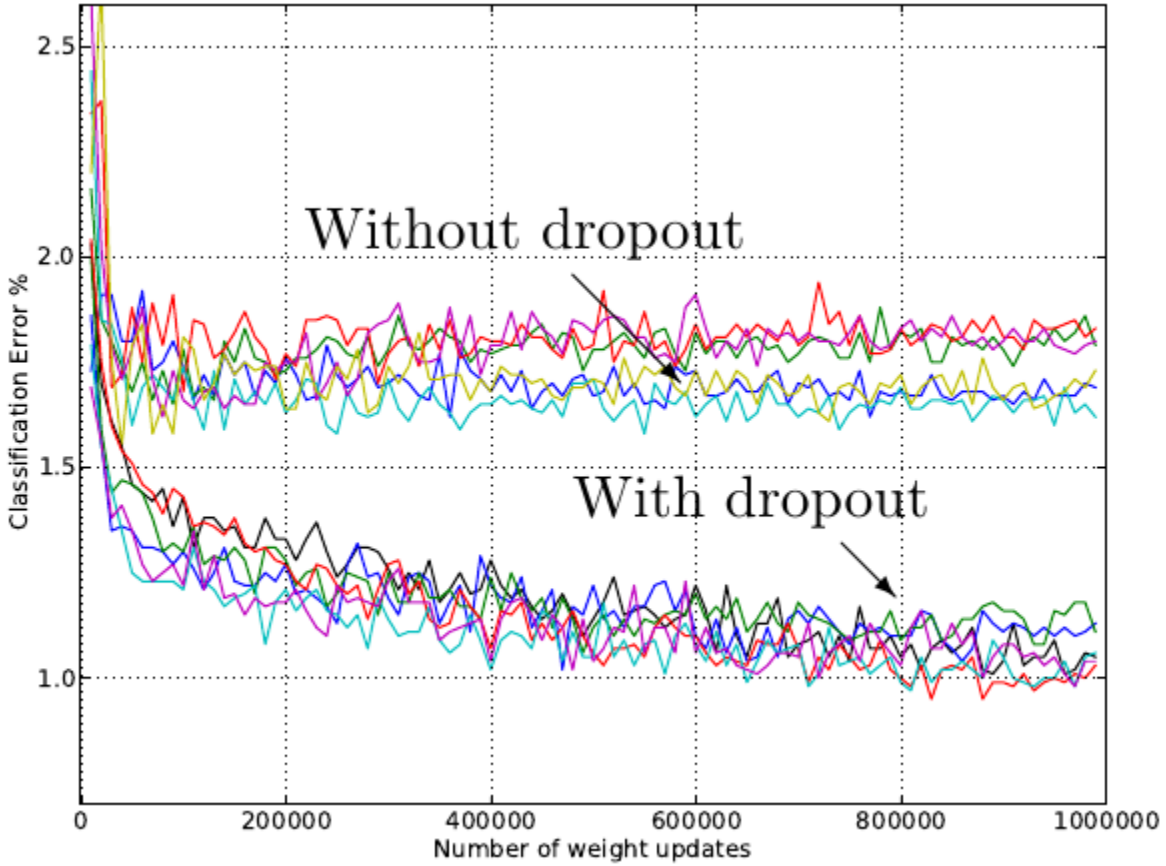


(a) At training time (b) At test time

Left: A unit at training time that is present with probability p and is connected to units in the next layer with weights $\mathbf{w}$. **Right**: At test time, the unit is always present and the weights are multiplied by p. The output at test time is same as the expected output at training time.

Srivastava, Nitish, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (2014): 1929-1958.

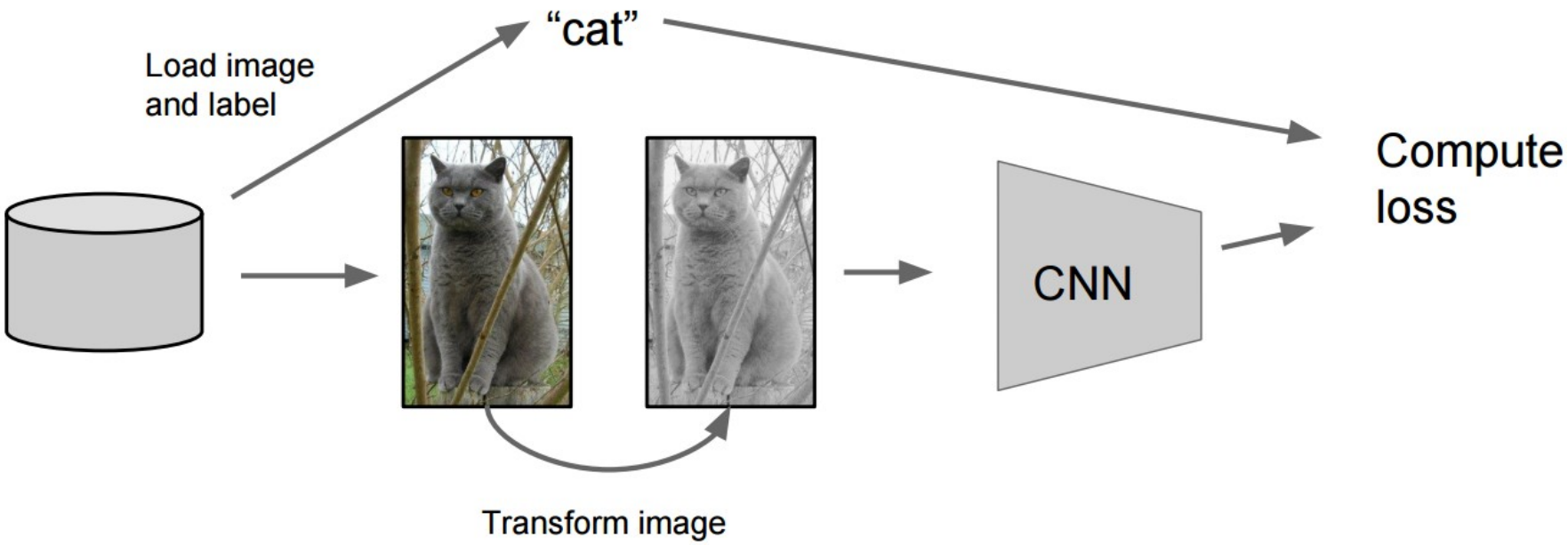
Dropout: Robustness



Test error for different architectures with and without dropout. The networks have 2 to 4 hidden layers each with 1024 to 2048 units.

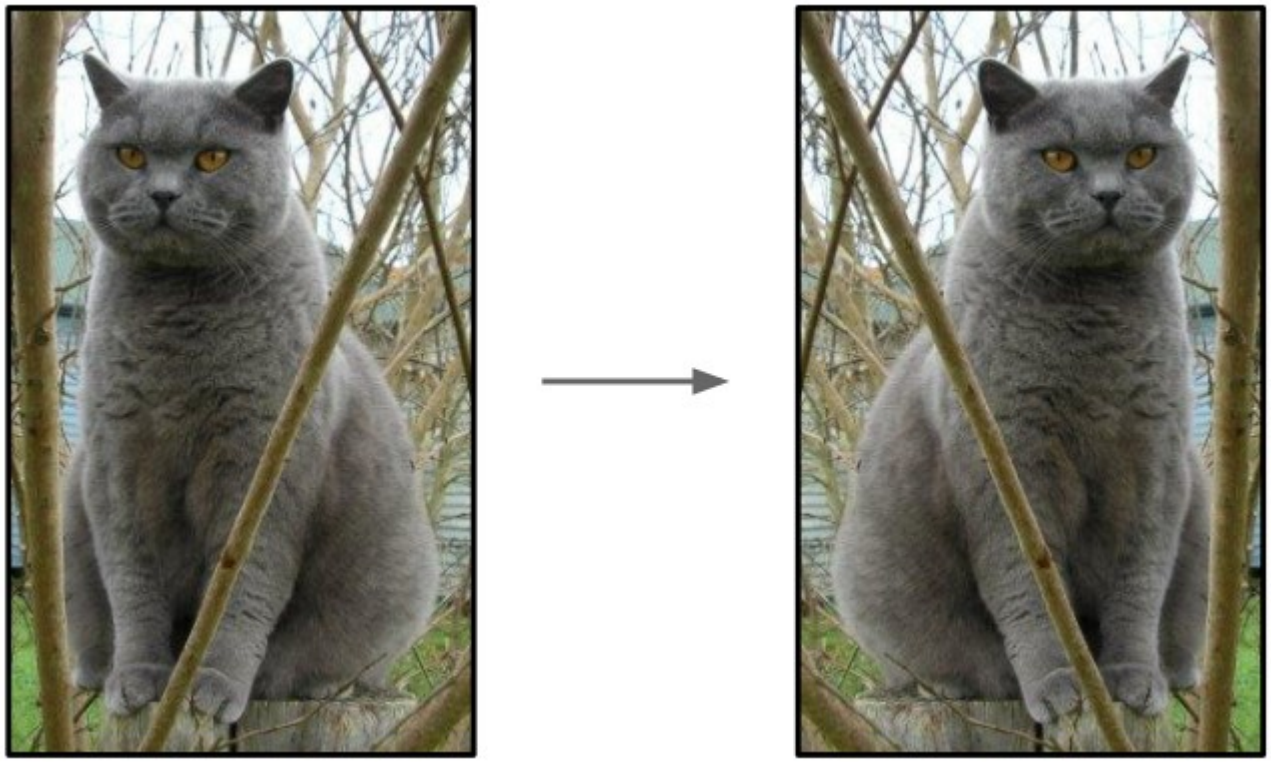
Srivastava, Nitish, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (2014): 1929-1958.

Data Augmentation



http://cs231n.stanford.edu/slides/winter1516_lecture11.pdf

Data Augmentation: Horizontal flips



http://cs231n.stanford.edu/slides/winter1516_lecture11.pdf

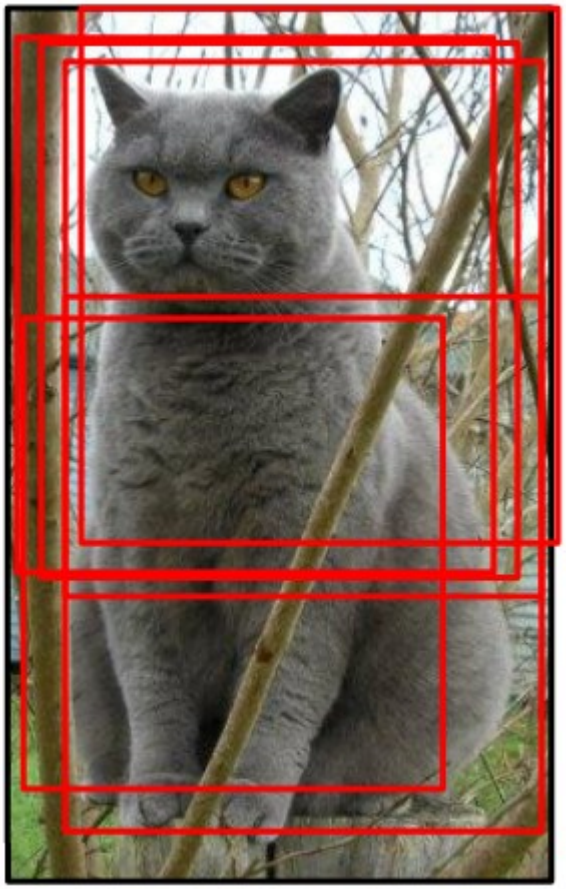
Data Augmentation: Random crops/scales

Training: sample random crops / scales
Por ejemplo [ResNet]:

1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224 x 224 patch

Testing: average a fixed set of crops
Por ejemplo [ResNet]

4. Resize image at 5 scales: {224, 256, 384, 480, 640}
5. For each size, use 10 224 x 224 crops: 4 corners + center, + flips



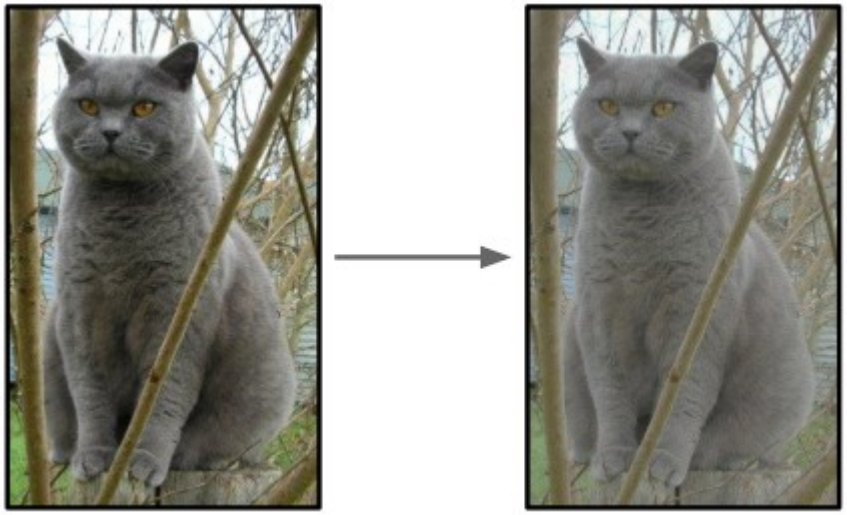
[ResNet] He, Kaiming, et al. "Deep residual learning for image recognition." arXiv preprint arXiv:1512.03385 (2015).

http://cs231n.stanford.edu/slides/winter1516_lecture11.pdf

Data Augmentation: Color jitter

Versión simple:

Alterar el contraste aleatoriamente



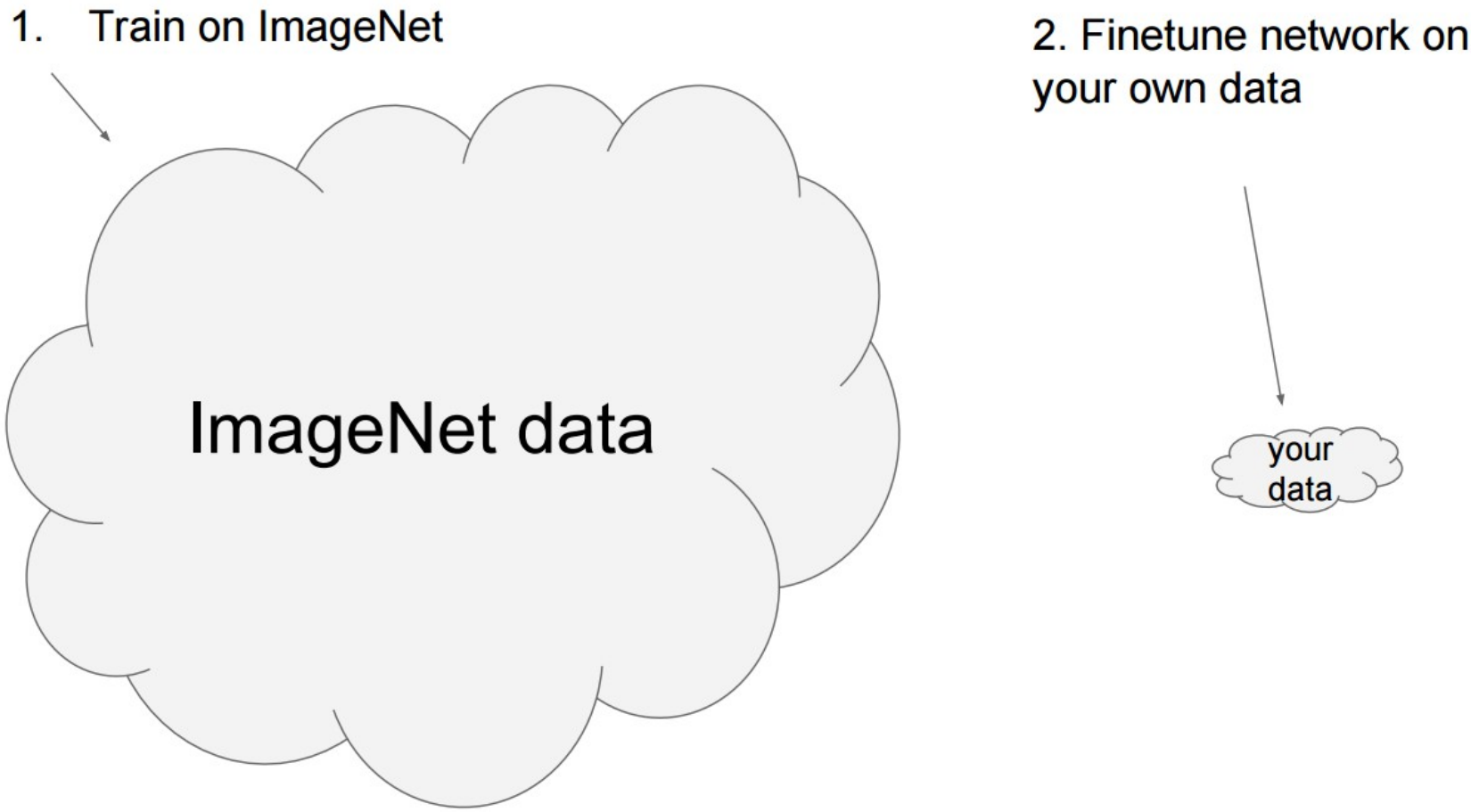
Versión compleja:

1. Aplicar PCA a todos los píxeles [R,G,B] del training set
2. Samplear un "desplazamiento de color" a lo largo de las direcciones principales
3. Aplicarle este desplazamiento a todos los píxeles de una imagen de entrenamiento.

http://cs231n.stanford.edu/slides/winter1516_lecture11.pdf

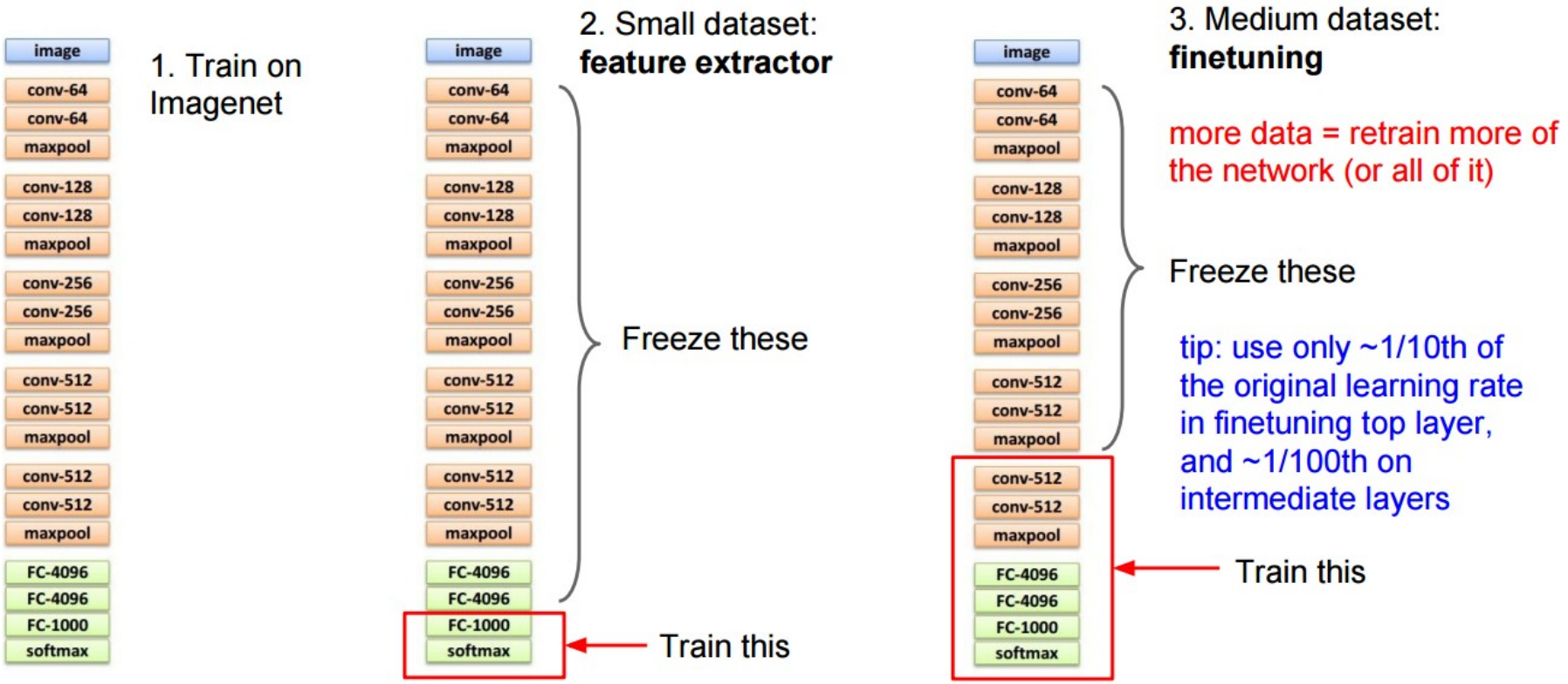
Transfer Learning

Pre entrenamiento sobre ImageNet



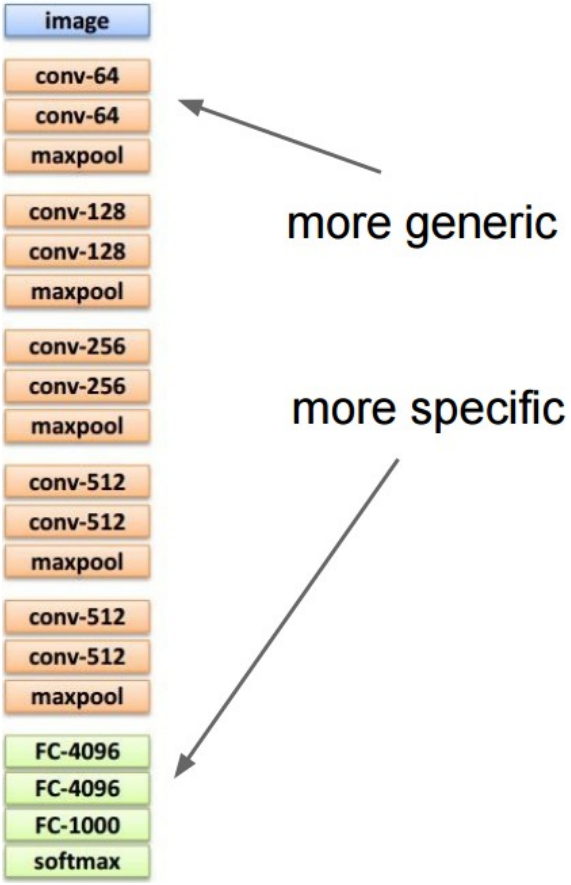
http://cs231n.stanford.edu/syllabus.html

Ejemplo con redes convolucionales



http://cs231n.stanford.edu/syllabus.html

Ejemplo con redes convolucionales



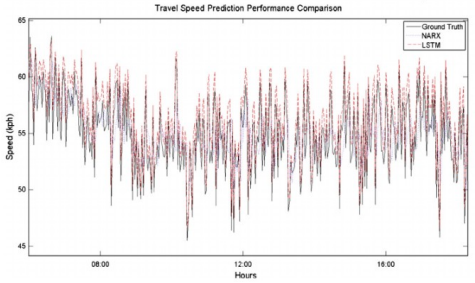
	very similar dataset	very different dataset
very little data	Use Linear Classifier on top layer	You're in trouble… Try linear classifier from different stages
quite a lot of data	Finetune a few layers	Finetune a larger number of layers

http://cs231n.stanford.edu/syllabus.html

Datos secuenciales

Datos Secuenciales

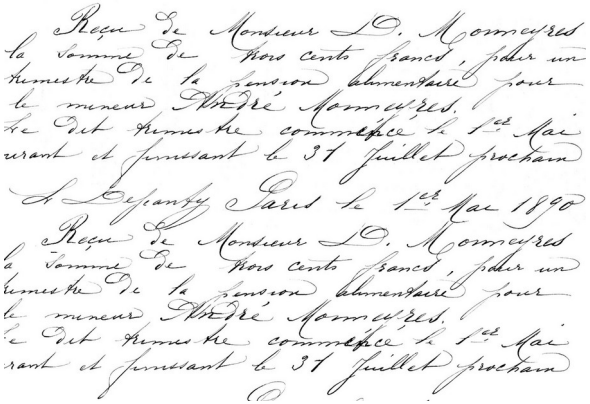
time series



text

ıa mirada panteísta donde un solo hombre inmortal es todos
ombres y a su vez ninguno. Y a partir de esta idea también pue
irmarse, como luego veremos, que un solo texto también
dos los textos. Según Borges este relato vendría a ser
osquejo de una ética para inmortales" y su tema "el efecto q
inmortalidad causaría en los hombres". Este efecto lo descri
orges a través del autor implícito del relato, el anticuario Jose
artaphilus, quien narra la vida del tribuno romano Mar
aminio Rufo. Así podremos presenciar en este relato la voz
ı hombre que fue todos y a la vez fue nadie, ya que fueron "
ılabras de otros […] la pobre limosna que le dejaron las hora
s siglos". El texto presente nos servirá para hacer una reflexi

handwriting



speech

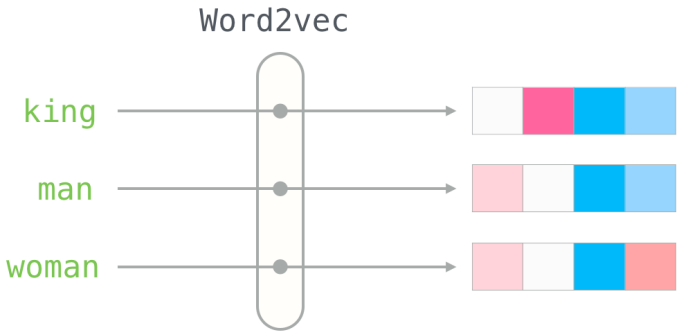
code

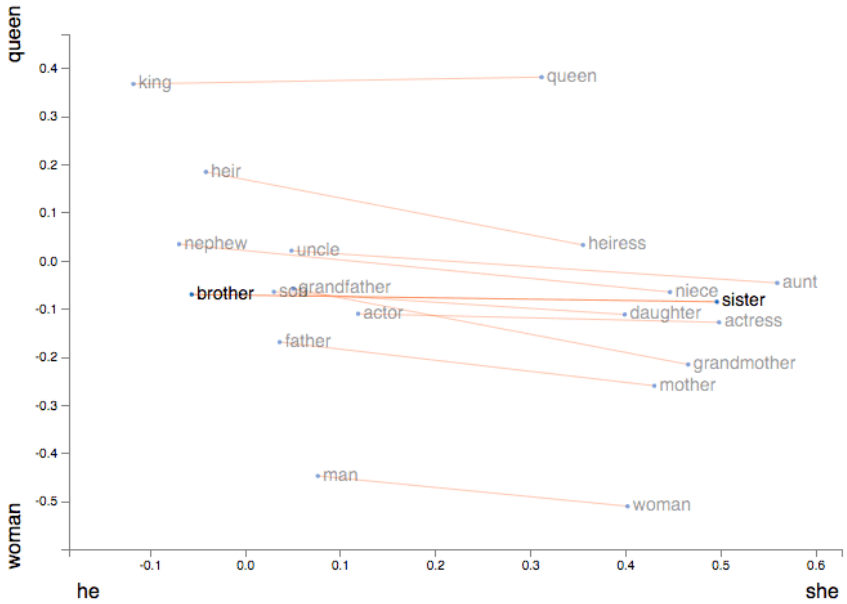
```
require_once('chorus/Shard.php');
Database::set_defaults(
  array('user' => 'tumblr3', 'password' => 'm3MpH1C0Koh39AQD83TFhsBPlOM1R
        'database' => 'tumblr3', 'write_lock_tables' => '*',
        'extended_log' => (idate('G') == 17 && intval(idate('i')) == 56

if ( __FILE__ == '/var/www/apps/tumblr/config/config.php' || __FILE__ == '/
  define('ENVIRONMENT', 'production');
  if (!defined('DEFAULT_DATABASE')) define('DEFAULT_DATABASE', 'primary');
  define('S3_BUCKET', 'data.tumblr.com');
  define('ENABLE_PANTHER', true);
  define('ENABLE_MEDIA_CDN', true);
  define('ASSETS_URL', (ENABLE_MEDIA_CDN && !(isset($_SERVER['HTTPS']) &&
  define('MEMCACHE_HOST', '10.252.0.68');
  define('MEMCACHE_VERSION_HOST', '10.252.0.67');
  define('VALIDATION_FAILURE_LOG', BASE_PATH . '/validate.log');
  define('REDIRECT_403_LOG', BASE_PATH . '/403.log');
  define('GOOGLE_API_KEY', (isset($_SERVER['HTTP_HOST']) && $_SERVER['HTTP
  'ABQIAAAAJlAd0HJn-kbPSqUsrS6CyhTpoeXstiwCMpsI5pU3slU-WDRPJxQts41ksQogKsy
  'ABQIAAAAJlAd0HJn-kbPSqUsrS6CyhTRJXjjauvD2gSXXVziOjeBJgmKOBTPW-8l5i5Pbk9
  Database::add('primary', array('host' => '192.168.200.142'));
  Database::add('db-tumblelogs', array('host' => '192.168.200.103'));
```

stock market

Texto: Word embedding

El texto no se puede usar directamente como las imágenes. Se necesita un espacio vectorial donde palabras con sentido similar tengan una representación similar.





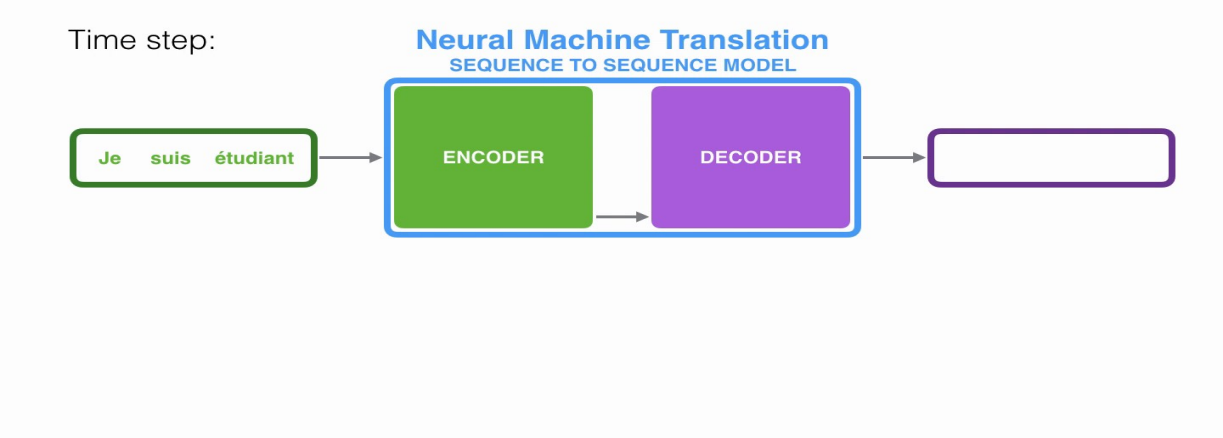
Learned: word2vec

https://machinelearningmastery.com/what-are-word-embeddings/

Aprendiendo una secuencia

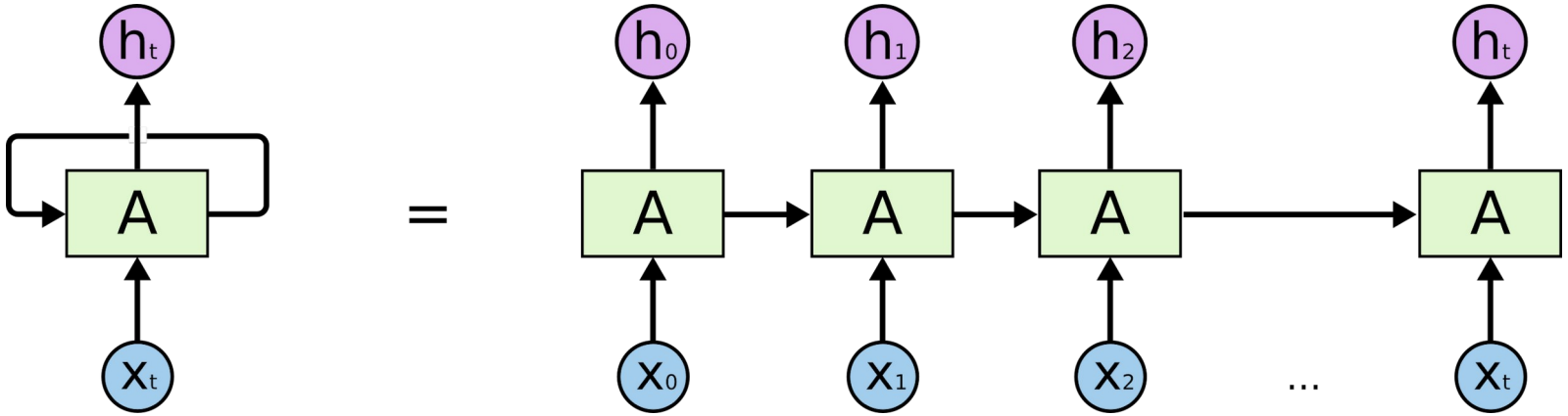
Para aprender una relación de secuencia-a-secuencia se usa un sistema en dos partes:

- Un encoder que traduce la secuencia a un estado interno que se aprende
- Un decoder que toma el estado interno y lo traduce a otra secuencia



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Recurrent Neural Networks



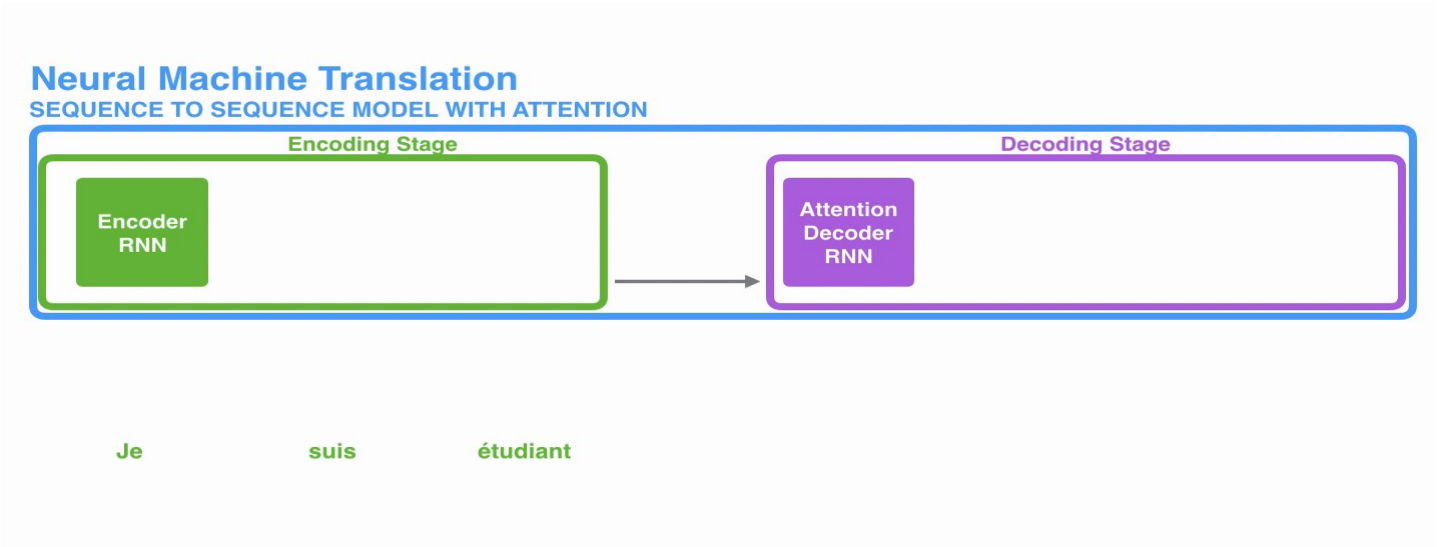
Neural Network with a loop

Unfolded Computational Graph

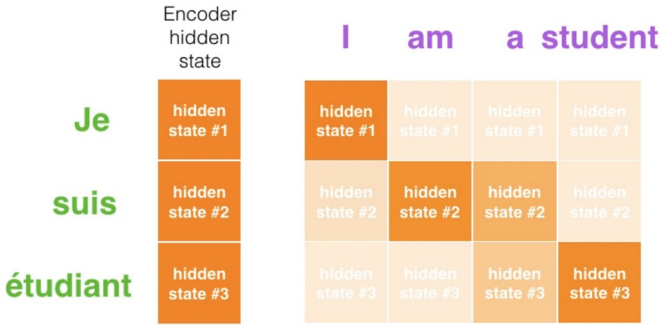
Problems: long-term dependencies + sequential learning

Christopher Olah's blog http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long term dependencies: Attention!

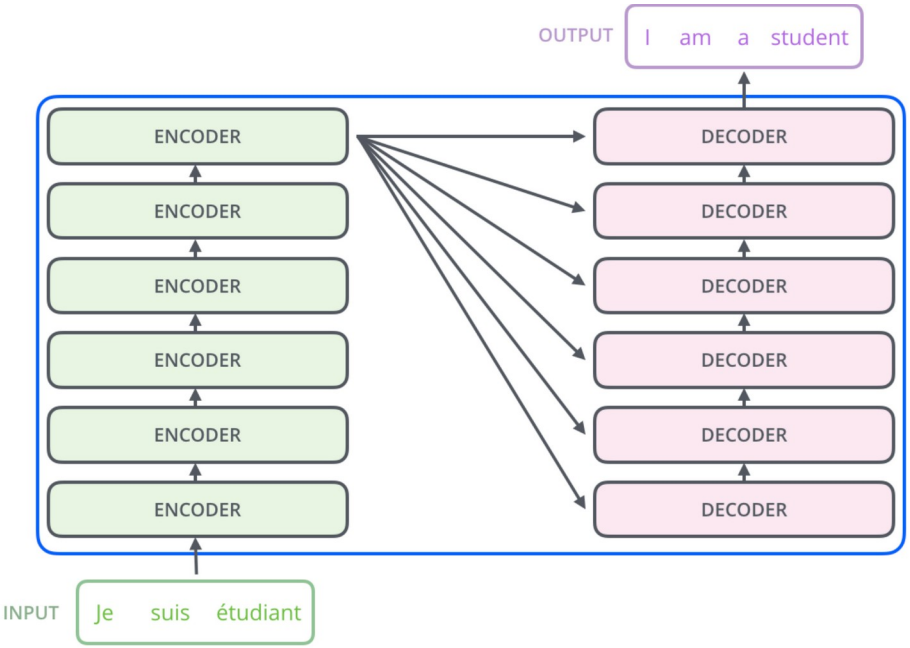


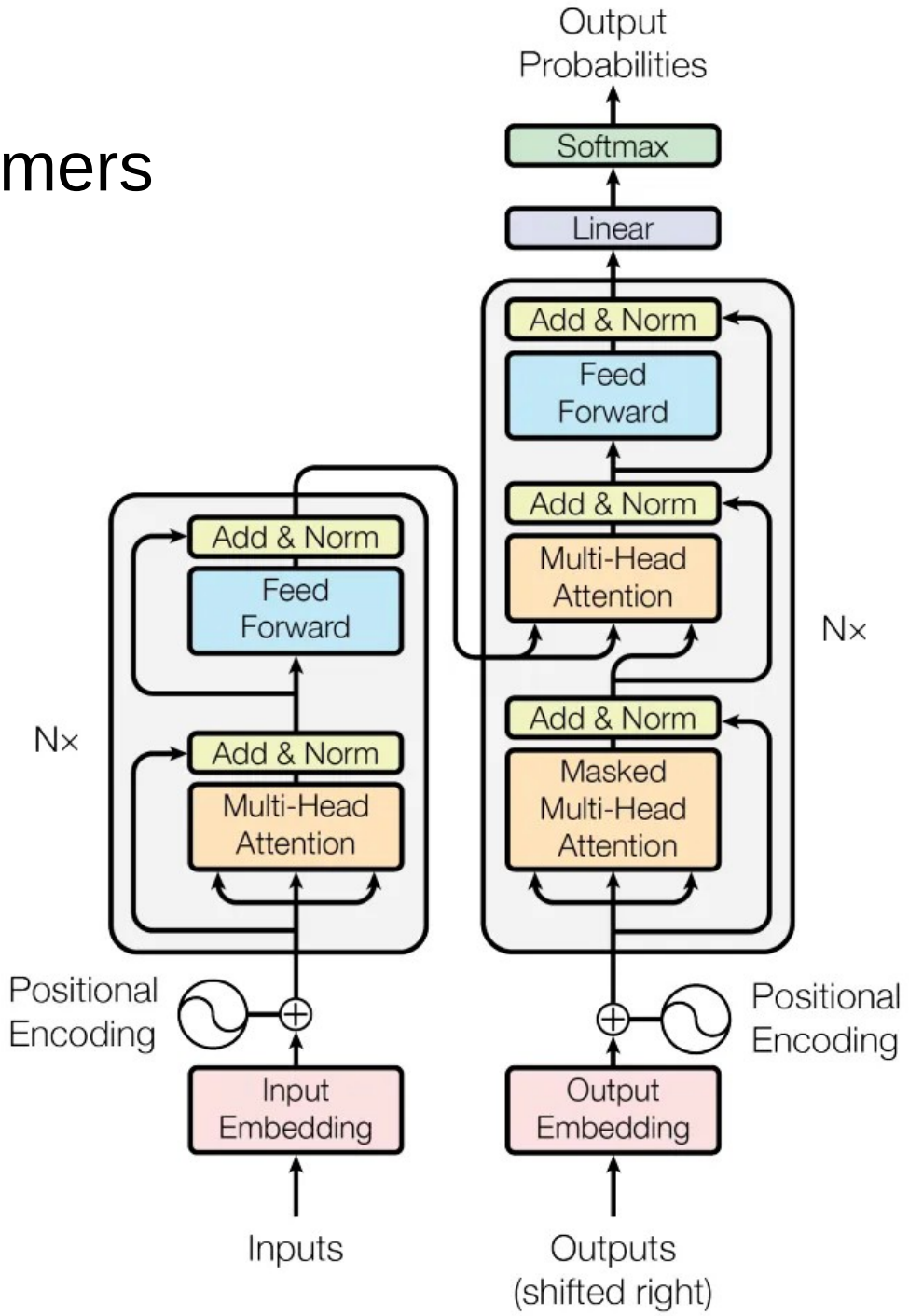
Permitimos al decoder ver todos los estados y elegir la información que usa en cada paso



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

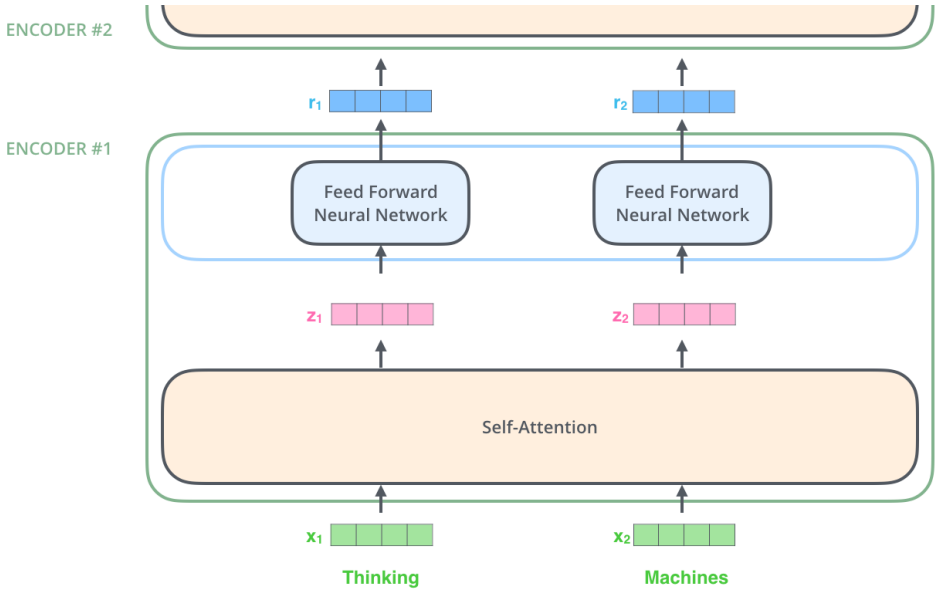
Parallel learning: Transformers



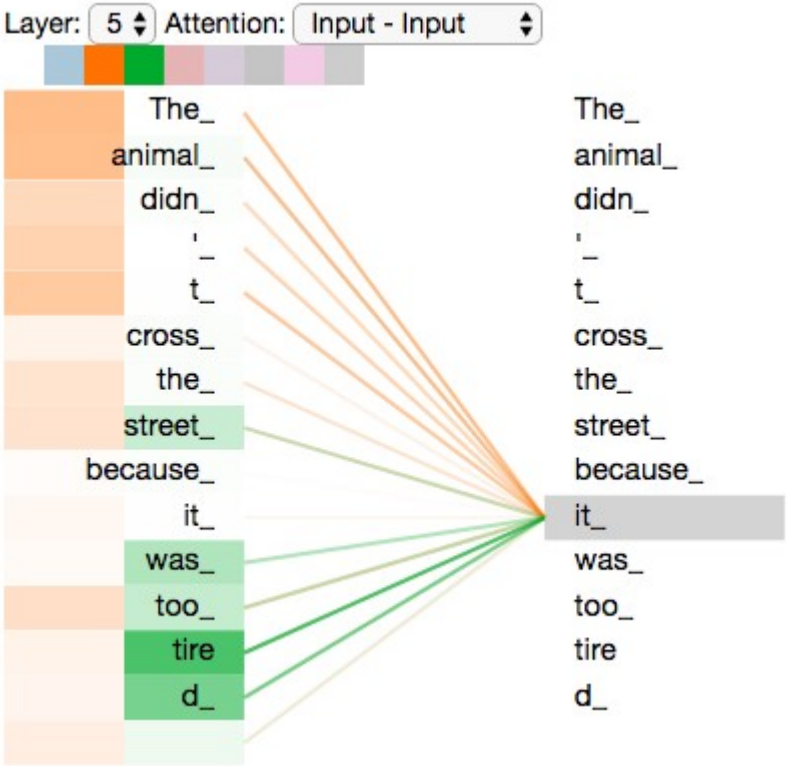


https://jalammar.github.io/illustrated-transformer/

Transformers: key points



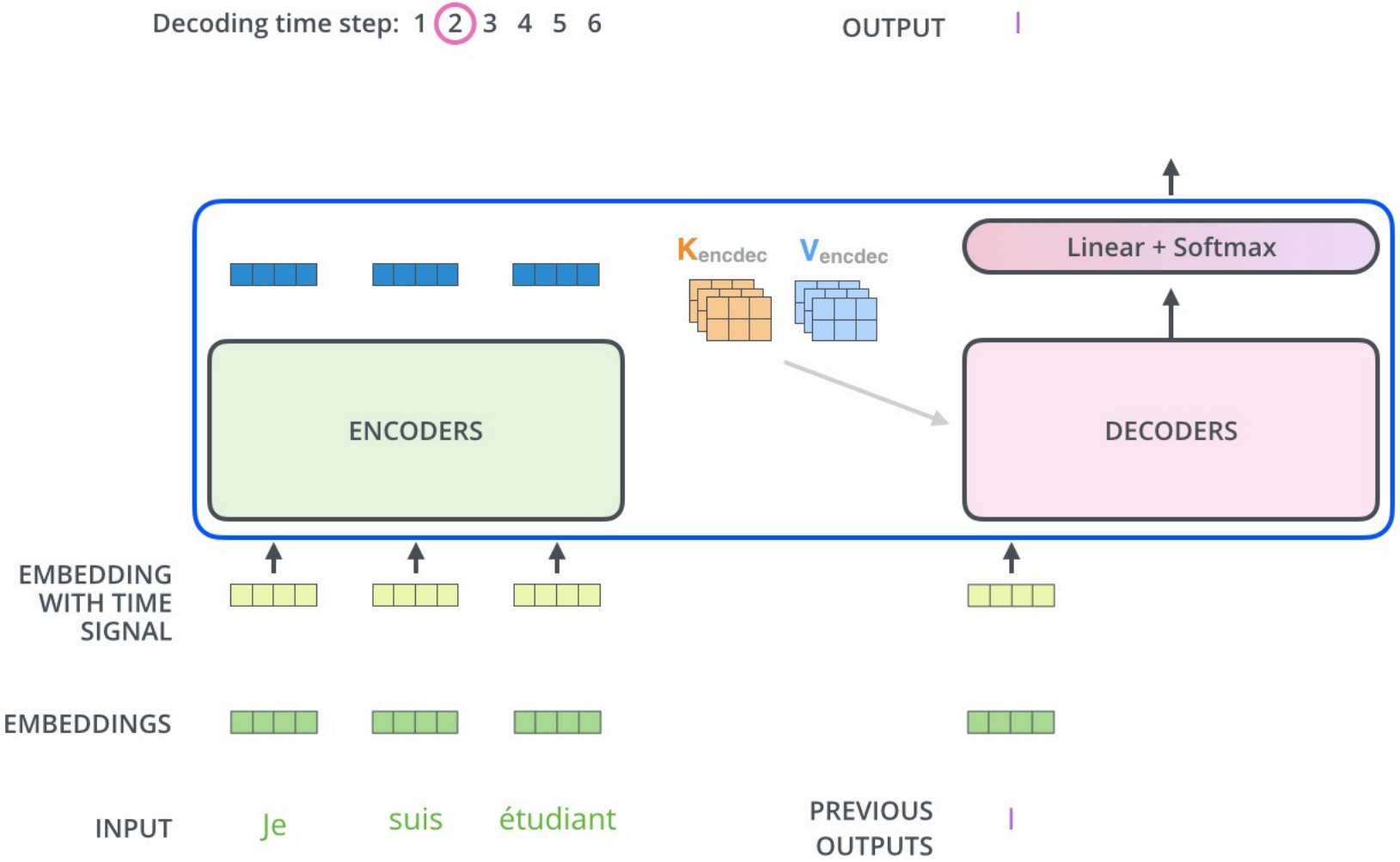
Multiple layers of encoders



Multiple attention heads

https://jalammar.github.io/illustrated-transformer/

Transformers: decoding



https://jalammar.github.io/illustrated-transformer/

Transformers: images

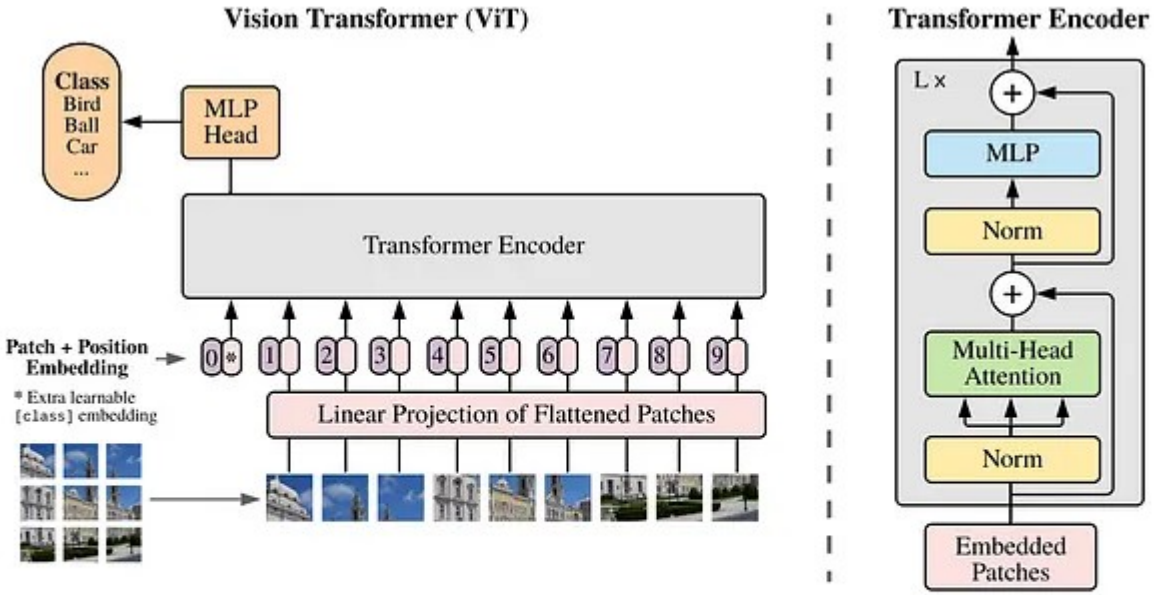


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale."