



Ph.D. Santiago Echeverri Arteaga

Aprendizaje por transferencia

Física Computacional 2

Transfer learning be like



Aprendizaje por transferencia

¿Qué es y para qué lo usaremos?

- En una red neuronal las primeras capas son las más difíciles de entrenar (Por gradientes desvanecidos)
- Ajustar las últimas capas bien tiene más relevancia en el resultado (Las primeras capas son las menos especializadas)
- Para un buen entrenamiento se necesita un conjunto de datos muy grande
- **Idea:** Guardar las primeras capas de un modelo pre-entrenado y re-entrenar las últimas capas para determinada aplicación (Ajuste fino)

A tener en cuenta para un ajuste fino

- Si los datos son similares a los de la red ore-entrenada, se necesita menos ajuste fino
- Entre más datos se tengan, más se beneficiará la red del ajuste fino
- Se congelan las capas que no se van a ajustar

Arquitecturas famosas

LeNet-5 AlexNet VGG GoogleLeNet InceptionV3 ResNet

...

LeNet-5

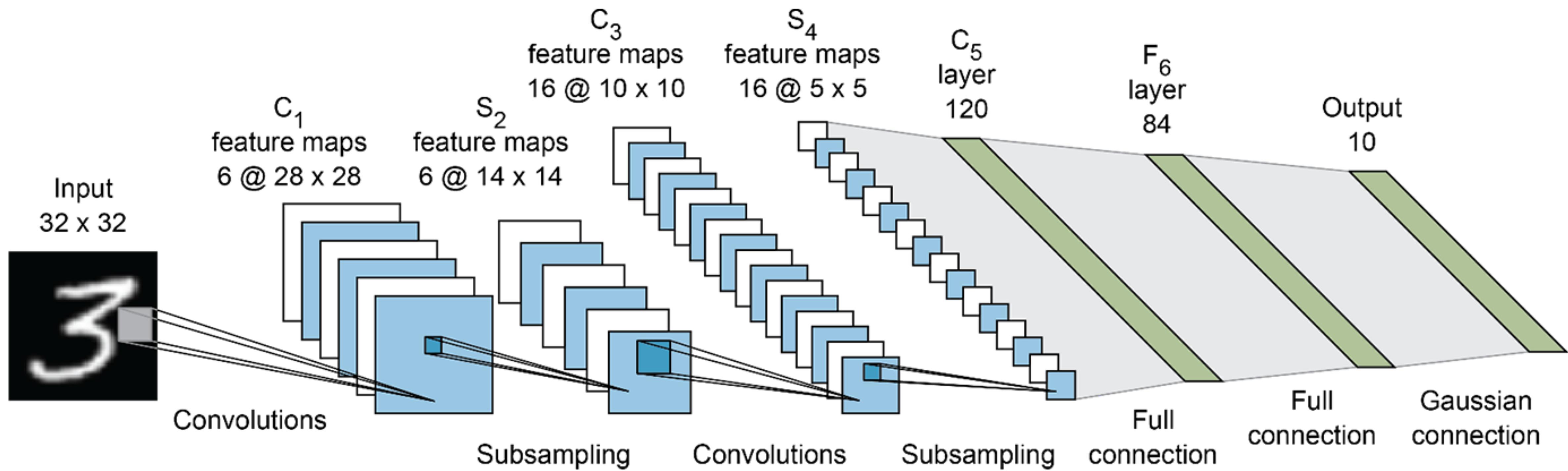
MINST

- Yann LeCun, Leon Bottou, Yoshua Bengio and Patrick Haffner en Bell Labs (1989) con 61706 pesos entrenados

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X			X	X	X		X	X	X	X	X	X	X	X	X
1	X	X			X	X	X		X	X	X	X	X	X	X	X
2	X	X	X			X	X	X		X	X	X	X	X	X	X
3		X	X	X		X	X	X	X		X	X	X	X	X	X
4			X	X	X		X	X	X	X		X	X	X	X	X
5				X	X	X		X	X	X	X		X	X	X	X

TABLE I
EACH COLUMN INDICATES WHICH FEATURE MAP IN S2 ARE COMBINED
BY THE UNITS IN A PARTICULAR FEATURE MAP OF C3.



IMAGENET

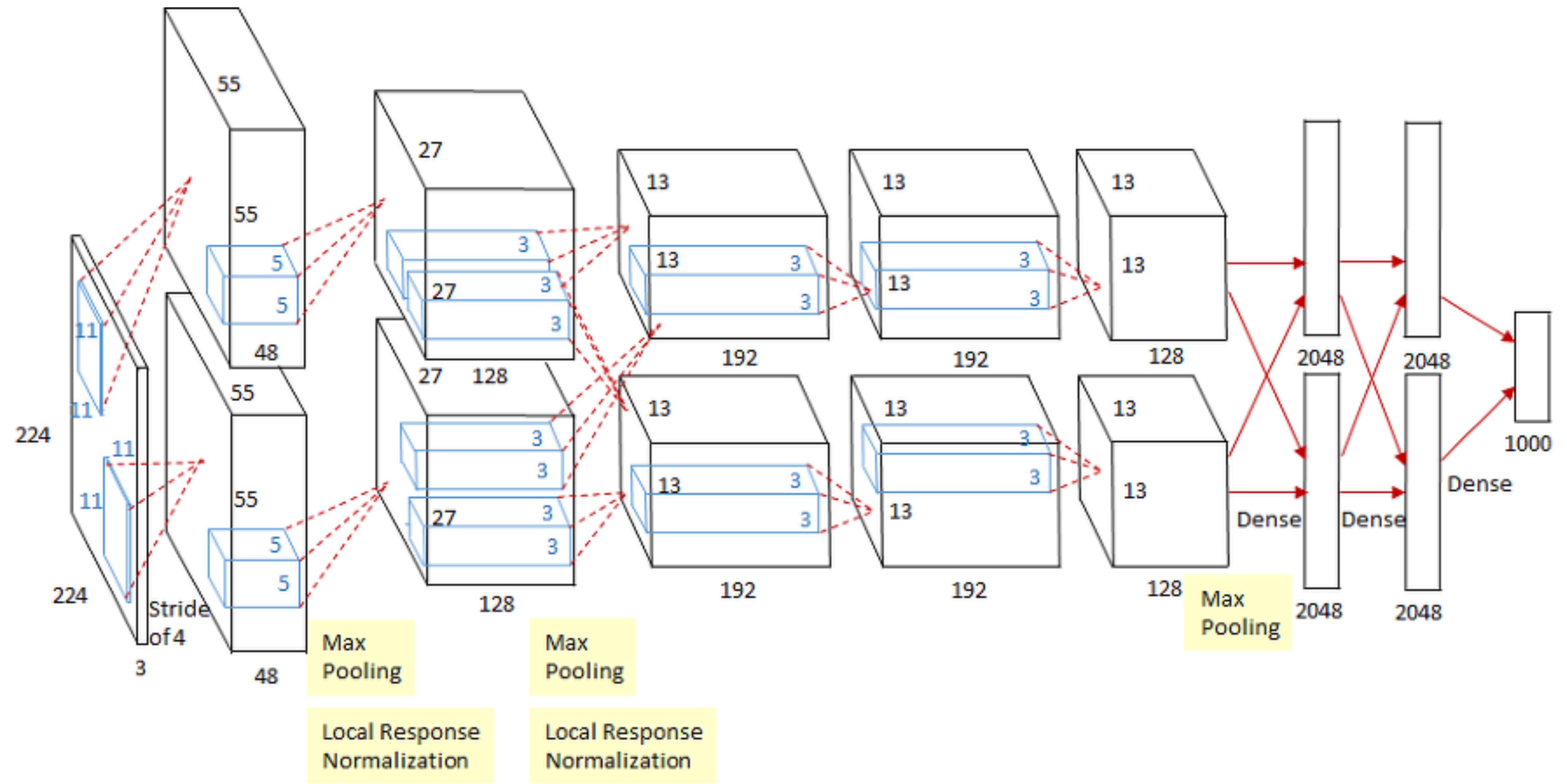
Lo que consiguió LeNet-5

- El éxito de LeNet-5 con el set de datos MNIST animó a la investigadora de Stanford University Fei-Fei Li a proponer la creación de un banco de datos de imágenes de todas clases, que pudieran servir de base para el entrenamiento de redes neuronales especializadas en el reconocimiento visual: ImageNet
- En ImageNet operarios humanos retribuidos etiquetaban manualmente imágenes, que luego se agrupaban en conjuntos semánticos. Cada conjunto semántico se integraba por aquellas imágenes cuyas etiquetas tenían significados semejantes. En la actualidad, el archivo de ImageNet contiene 100.000 conjuntos semánticos (synsets) diferentes, dotados de 1.000 imágenes cada uno.

AlexNet

Arrasando en el ILSVRC (ImageNet Large Scale Visual Recognition Challenge)

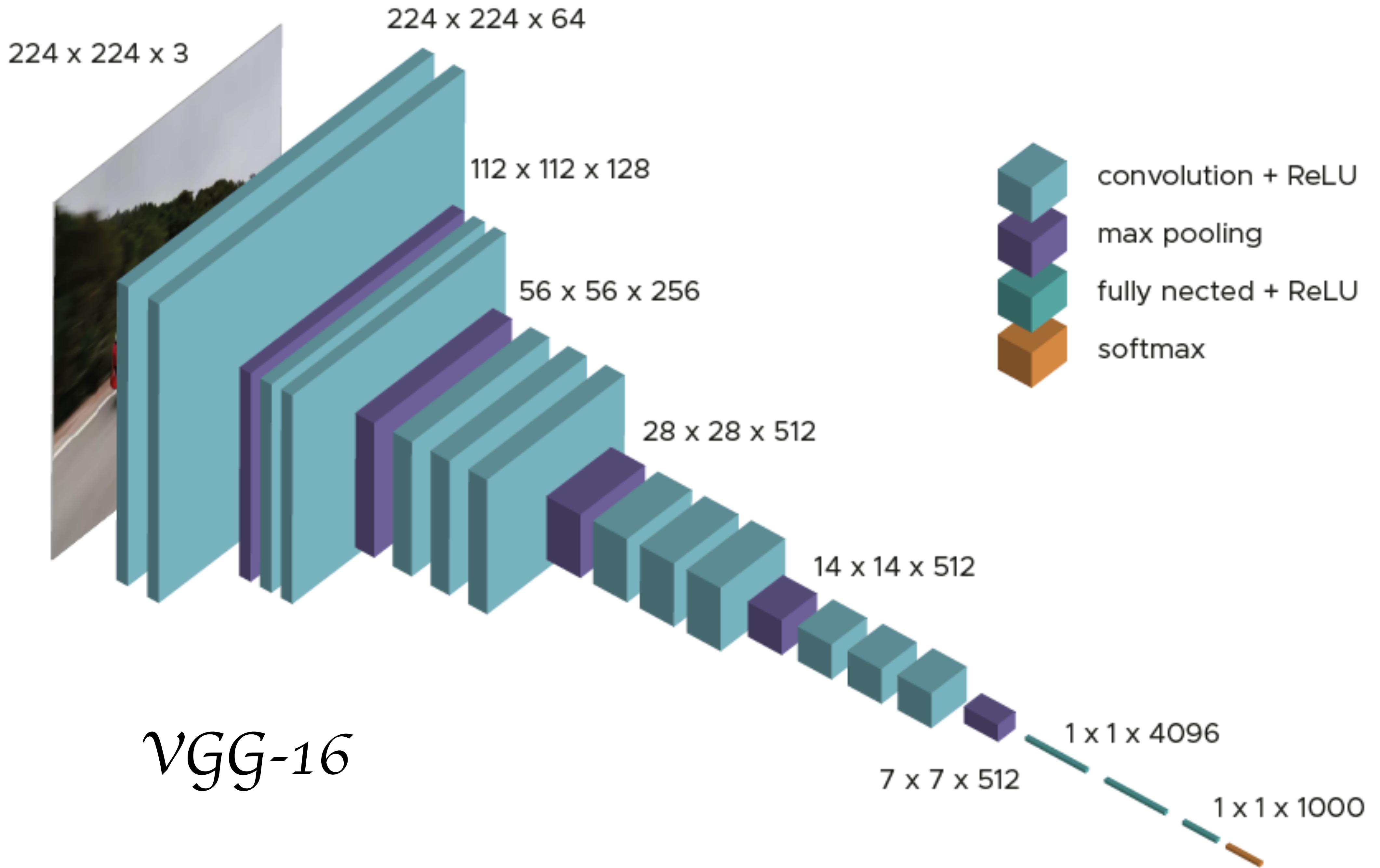
- El reto era clasificar imágenes dentro de 1000 categorías
- Top-5 error: La red propone cinco etiquetas diferentes a cada imagen. Si ninguna de ellas se corresponde con su descripción, se considera que se ha producido un fallo.
- El record que existía era un error Top-5 del 25%. Alex Krizhevsky y Geoffrey Hinton (dir) propusieron AlexNet que consiguió en el 2012 un error Top-5 del 15.4% en un set de datos de 1'200.000 imágenes
- Usaron función de activación ReLu y fue simulada por dos GPU
- Para analizar a AlexNet se introdujo la convolución inversa
- 60 millones de parámetros a entrenar (semanas entrenando)
- Aumentaron la cantidad de datos para evitar sobreajuste y mejorar predictibilidad con transformaciones sobre las imágenes



VGG

92.7% de precisión en ImageNet!

- K. Simonyan y A. Zisserman de Oxford
- Preprocesamiento: Restar a cada pixel el promedio de valor RGB (calculado en el conjunto de entrenamiento)
- Usa varias convoluciones consecutivas con kernels de 3x3 para asemejar una sola capa convolucional con mayor tamaño de kernel
- 3 capas 3x3 son 27 parámetros y tienen el mismo receptive field que una de 7x7 que tiene 49 parámetros (45% más parámetros)



THAT'S NOT ENOUGH

WE HAVE TO GO DEEPER

A movie poster for the film "Inception". The image depicts a massive skyscraper, the Tyrell Corporation building, which has collapsed and is falling into the ocean. The building is shown in various stages of collapse, with large amounts of dust and debris. In the foreground, several people are standing on a beach, looking out at the falling building. The sky is filled with dark, heavy clouds. The overall atmosphere is one of a catastrophic event.

INCEPTION

<https://arxiv.org/pdf/1409.4842v1.pdf>

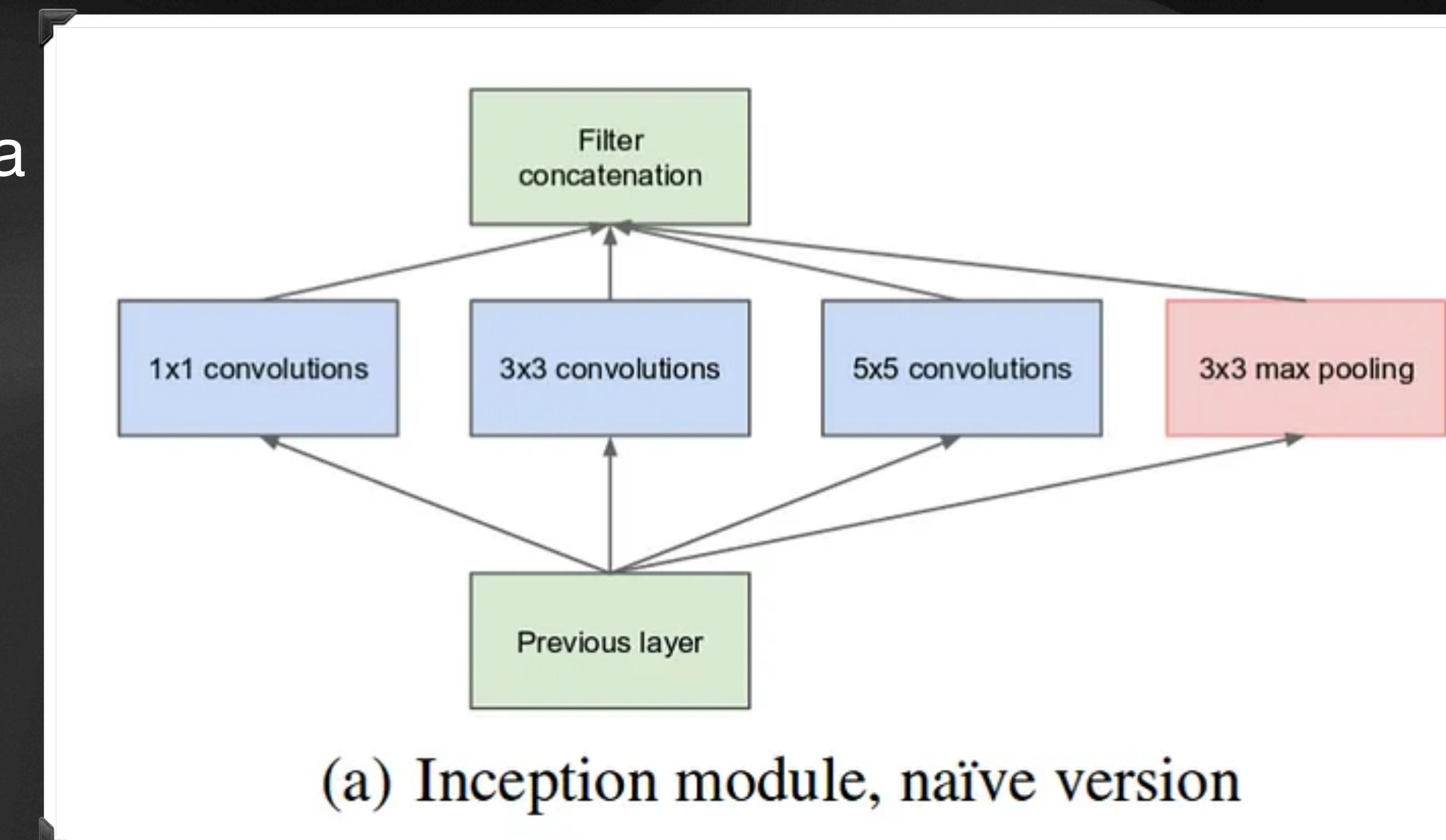
<https://arxiv.org/pdf/1512.00567v3.pdf>

<https://arxiv.org/pdf/1602.07261.pdf>

Inception



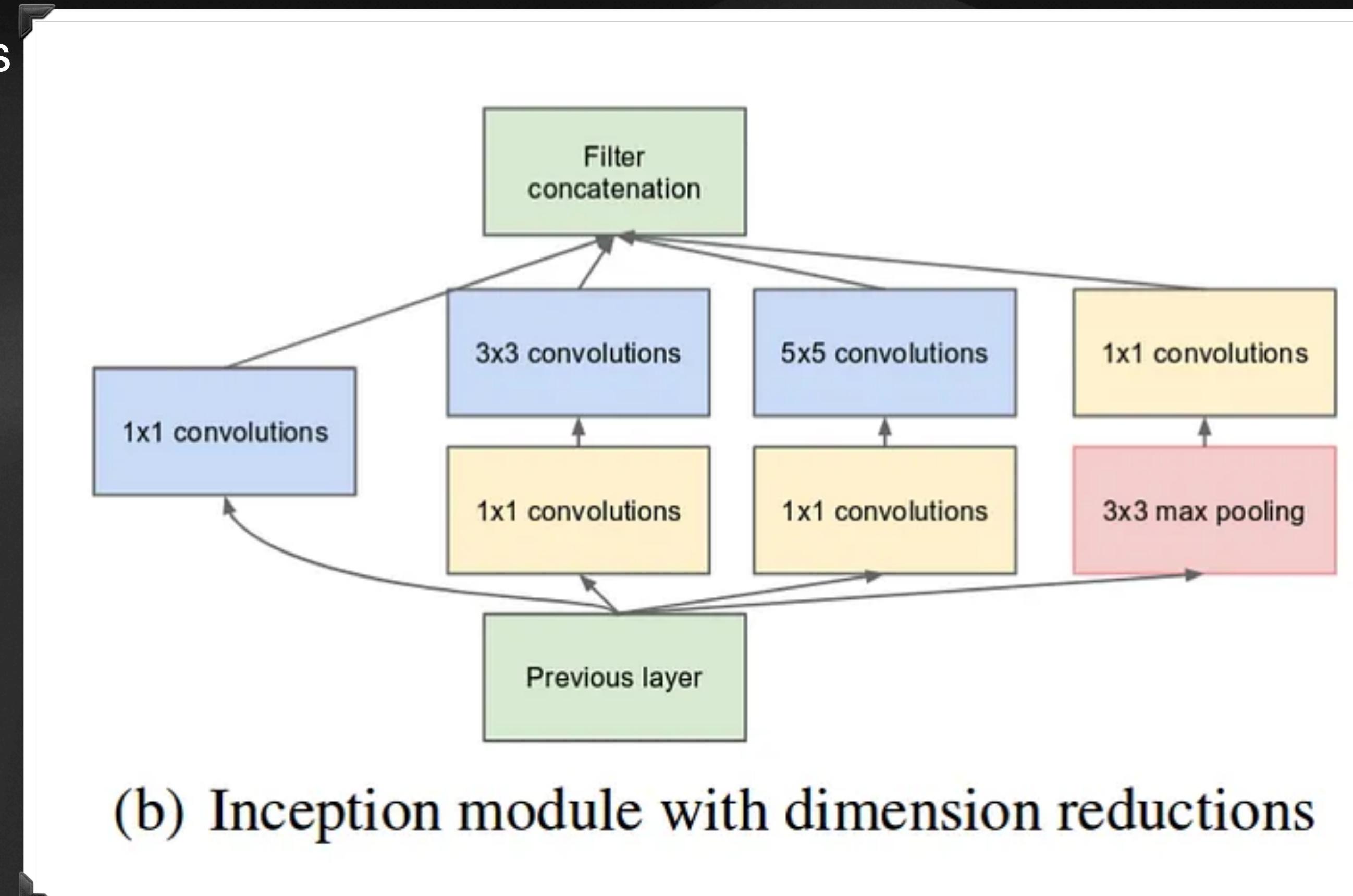
- El tamaño del kernel adecuado es fundamental. Un kernel más grande se prefiere si la información está distribuida
- Las redes muy profundas son susceptibles al sobreajuste
- Convoluciones son “costosas” computacionalmente
- **Solución:** Múltiples filtros operando en el mismo nivel (Wider than deeper)

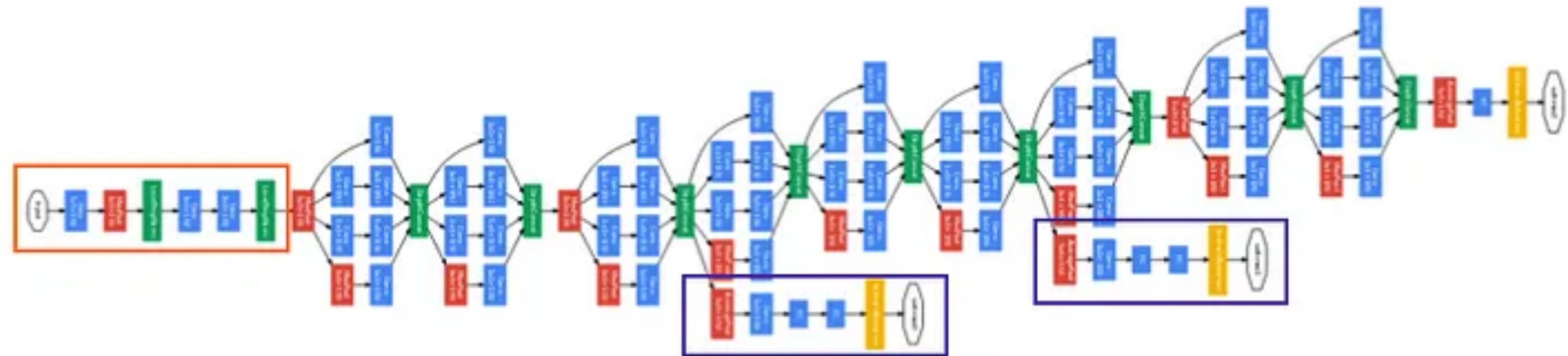


Inception



- Un filtro 1x1 sirve para “promediar” la salida de los filtros de entrada y así reducir la dimensionalidad (aumentando la eficiencia computacional)
- Inception v1 es llamado también GoogleLeNet
- Tiene 9 módulos de inception linealmente ensamblados. (27 capas ocultas incluyendo las pooling)
- Usa average pooling
- Introduce **durante el entrenamiento** clasificadores auxiliares que contribuyen al Loss
$$Loss = Loss_0 + 0.3L_1 + 0.3L_2$$
- Solo 5 millones de parámetros!





Inception V2

Cambios en el módulo de inception

Figure 5

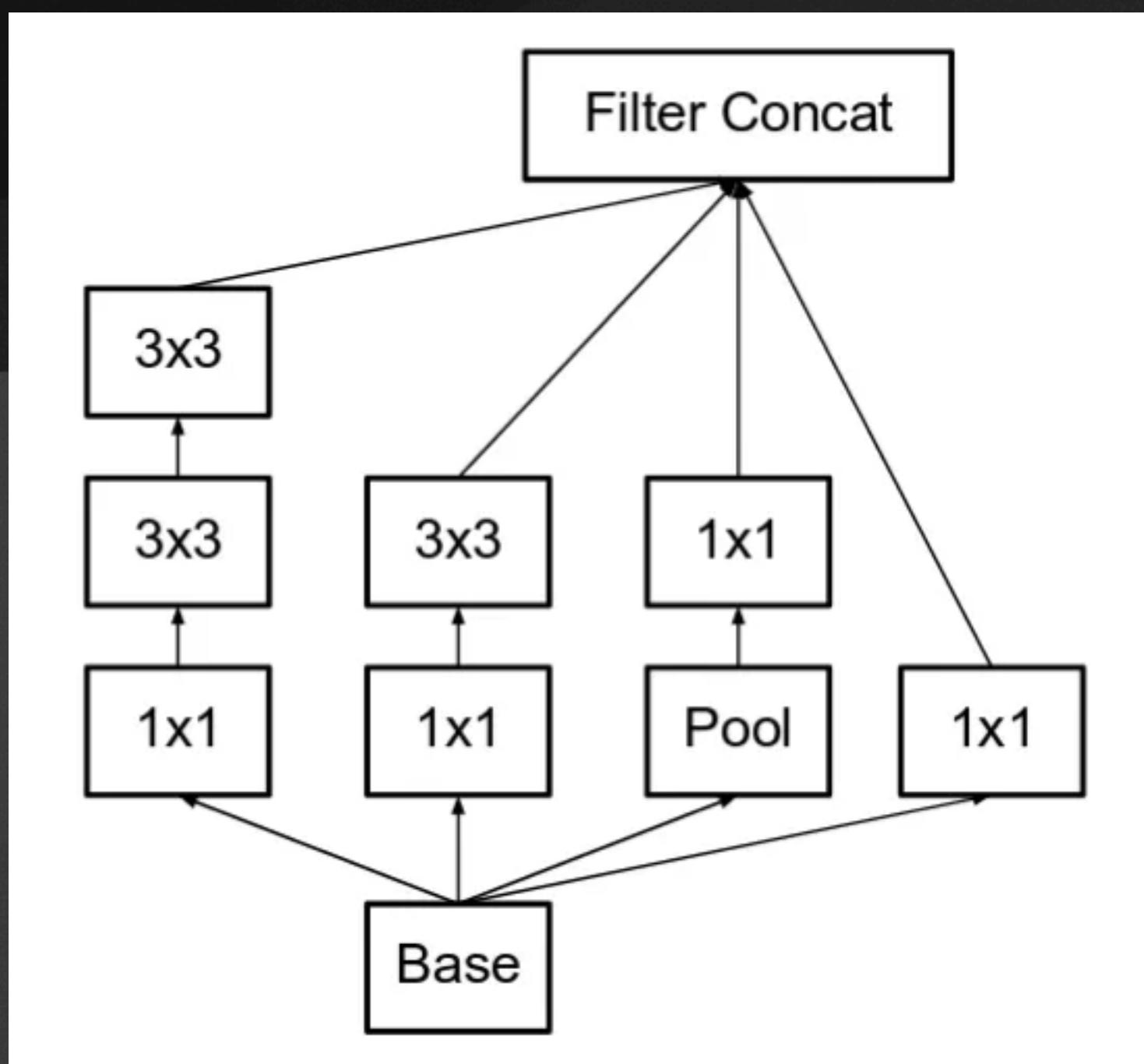


Figure 6

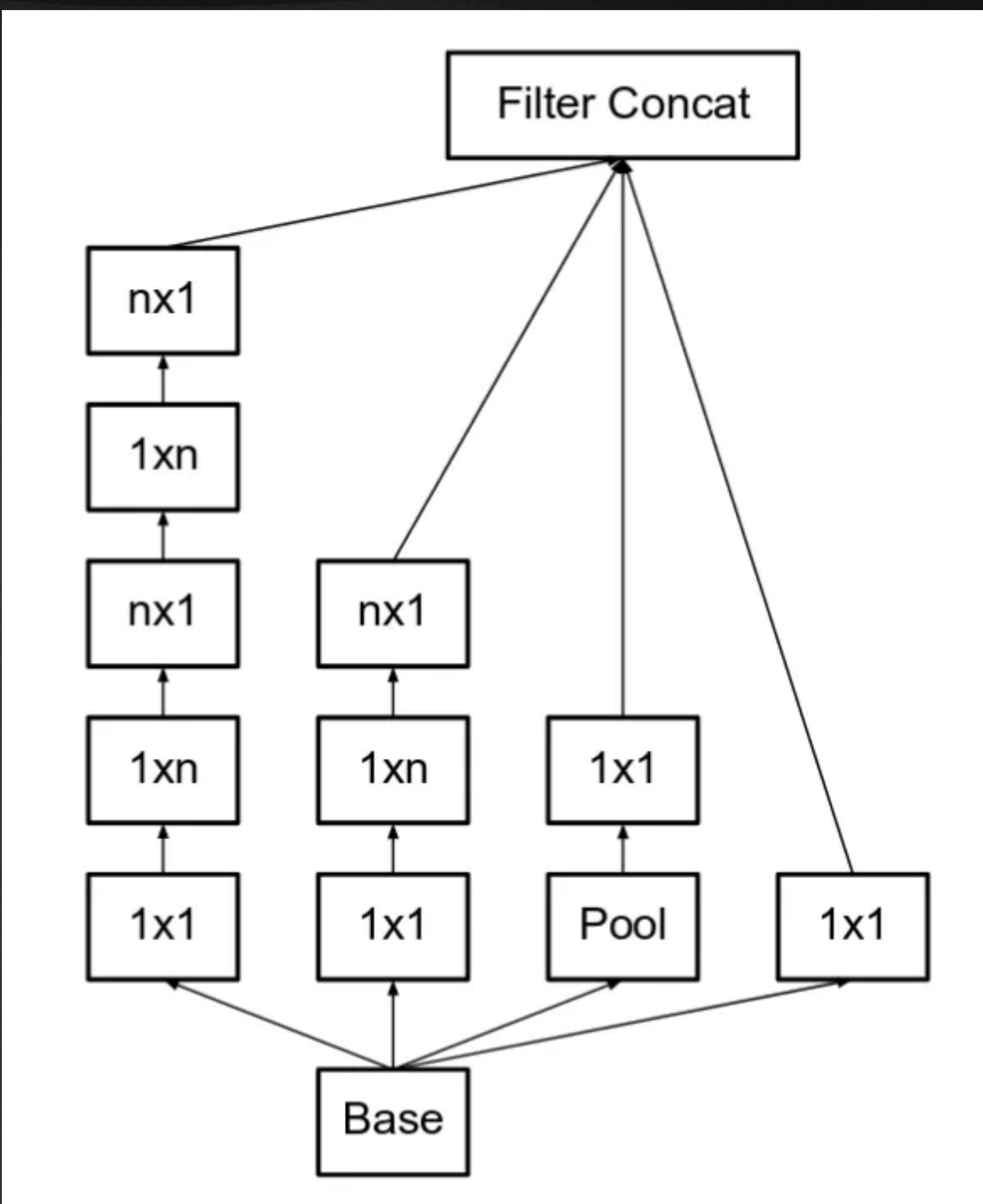
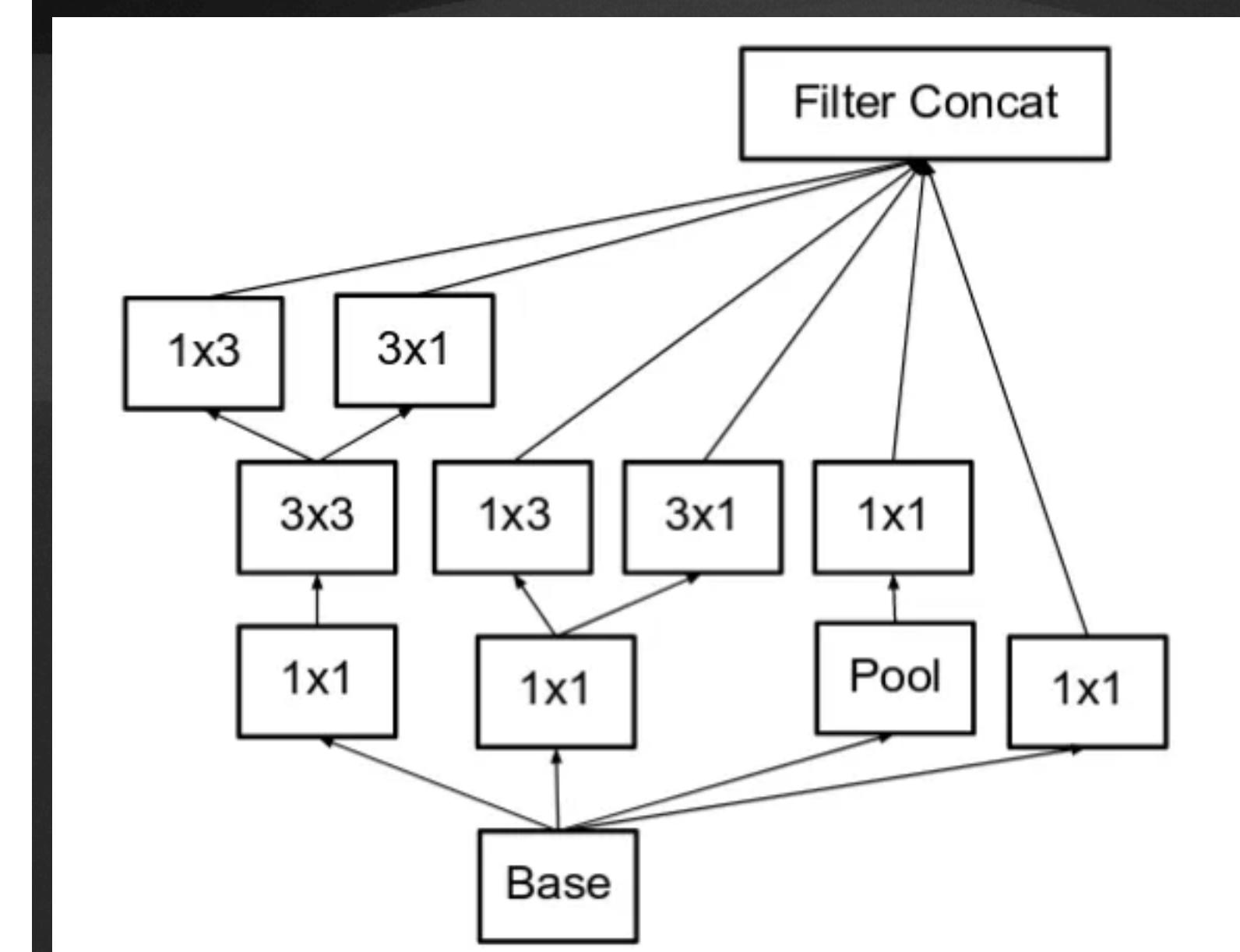


Figure 7



Inception V2

type	patch size/stride or remarks	input size
conv	$3 \times 3 / 2$	$299 \times 299 \times 3$
conv	$3 \times 3 / 1$	$149 \times 149 \times 32$
conv padded	$3 \times 3 / 1$	$147 \times 147 \times 32$
pool	$3 \times 3 / 2$	$147 \times 147 \times 64$
conv	$3 \times 3 / 1$	$73 \times 73 \times 64$
conv	$3 \times 3 / 2$	$71 \times 71 \times 80$
conv	$3 \times 3 / 1$	$35 \times 35 \times 192$
$3 \times$ Inception	As in figure 5	$35 \times 35 \times 288$
$5 \times$ Inception	As in figure 6	$17 \times 17 \times 768$
$2 \times$ Inception	As in figure 7	$8 \times 8 \times 1280$
pool	8×8	$8 \times 8 \times 2048$
linear	logits	$1 \times 1 \times 2048$
softmax	classifier	$1 \times 1 \times 1000$

Inception everywhere

[https://towardsdatascience.com/a-simple-guide-to-the-VERSIONS-OF-THE-INCEPTION-NETWORK-7fc52b863202](https://towardsdatascience.com/a-simple-guide-to-the VERSIONS OF THE INCEPTION NETWORK-7fc52b863202)

- Inception V3, V4 Inception ResNet V1, V2
- Uso de RMSProp optimizer, convoluciones 7x7 optimizadas, Batch Normalization, Componentes personalizados para regularización y cambios en la arquitectura (V3 y V4)
- Combinación con ResNet para Inception ResNet V1, V2

Google

how to train a resnet in pytorch



All Images Videos News Maps More

Settings Tools

About 815,000 results (0.39 seconds)

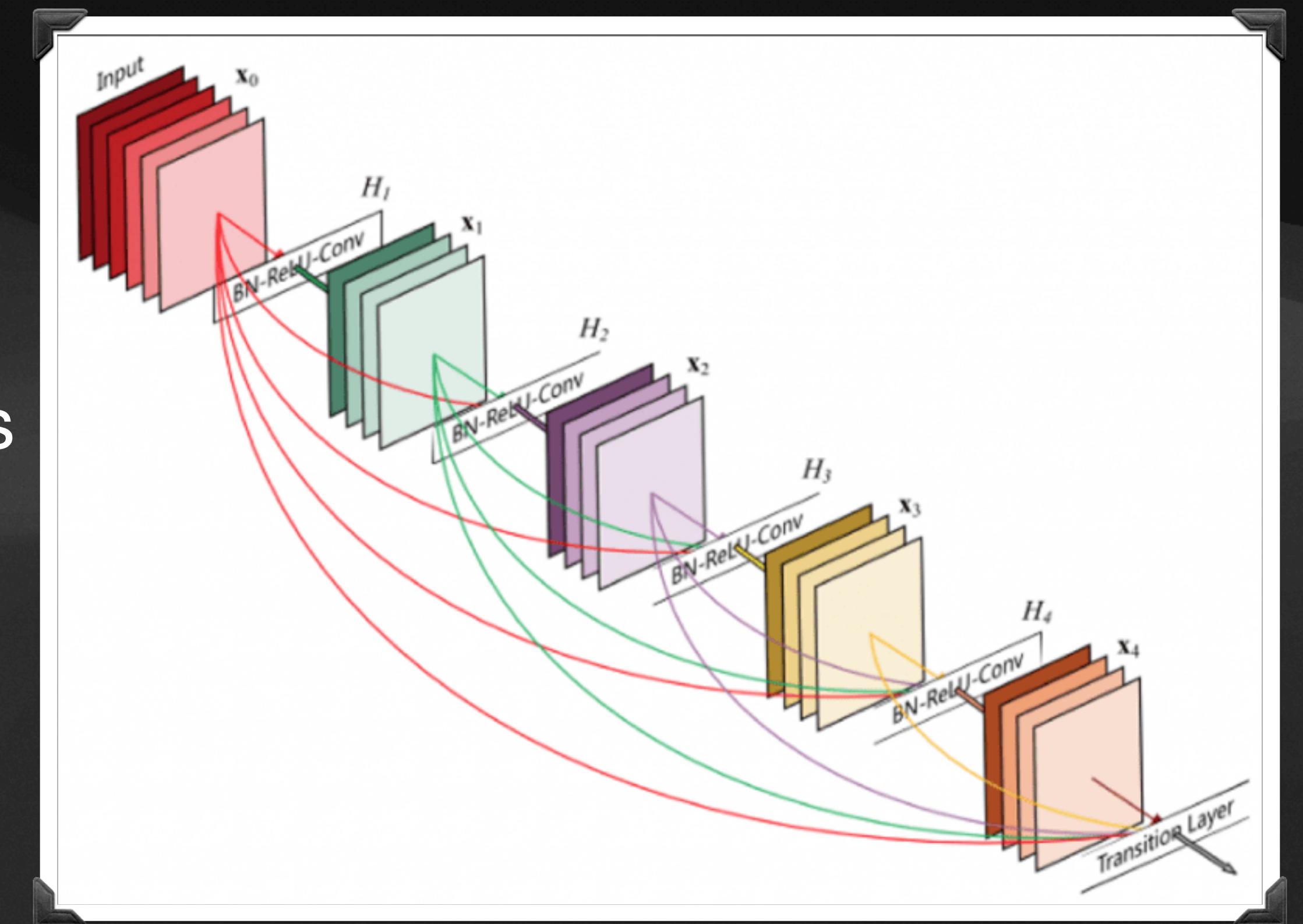
Did you mean: *how to train a resnet in tensorflow*



ResNet

Red neuronal residual

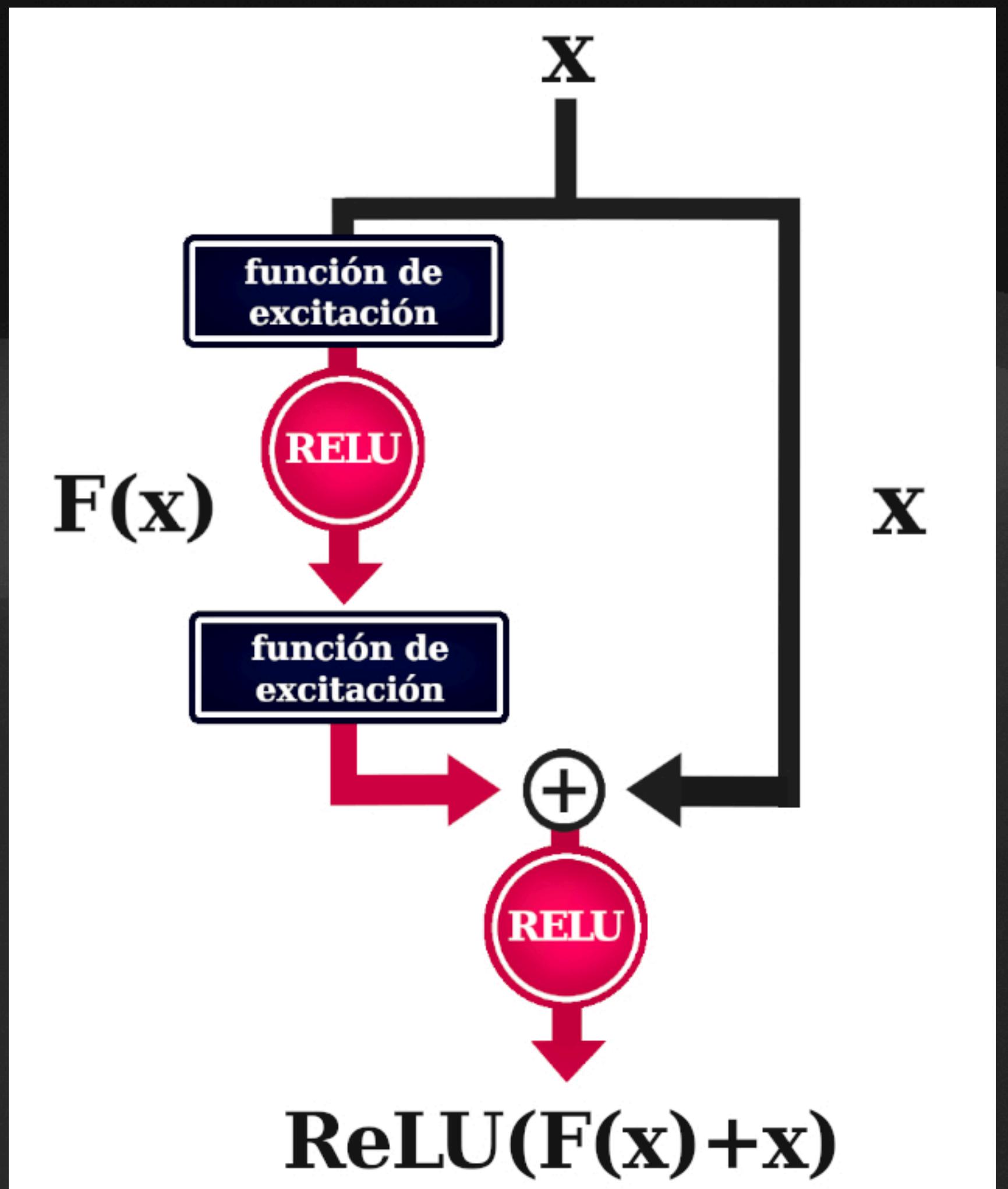
- Kaiming He et. al. (Microsoft Research)
“Deep Residual Learning for Image Recognition” (2016).
- Uso de atajos para moverse entre capas y solventar el error de las redes muy profundas
- Consideración ResNet: El mejor ajuste a múltiples capas es cercano a $\mathcal{F}(x) \sim \mathcal{F}(x) + x$
 - Se agregan conexiones de cortocircuito
 - Cuando se tienen múltiples saltos paralelos se llaman densnets



ResNet

Red neuronal residual

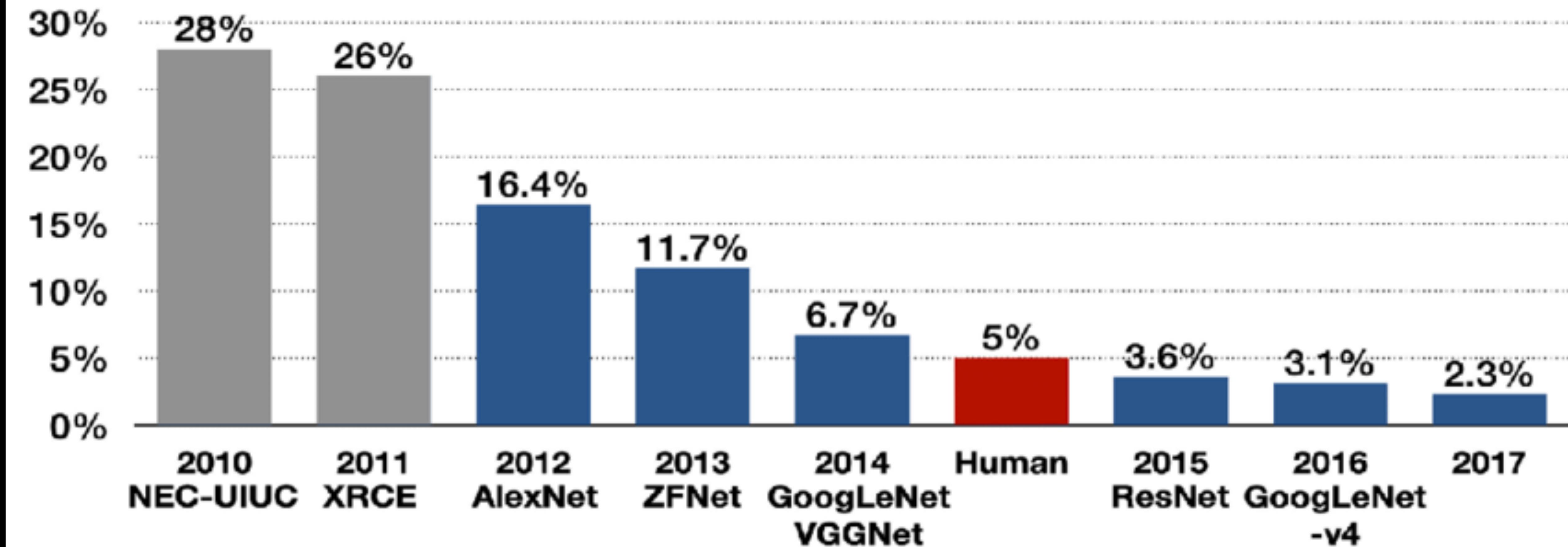
- Funciones de excitación pueden ser redes densas o convolucionales
- Bloque residual: $\mathcal{H}(x) = \mathcal{F}(x) + x$
- Si en las funciones de excitación se hace un cambio de dimensionalidad, se debe incorporar una matriz de proyección que asegure la misma modificación en la conexión de atajo
- El gradiente atraviesa las conexiones de atajo y por tanto se evita el problema de gradientes desvanecidos
- Resnet 50, 101, 152 superaba a VGG16 y GoogleLeNet en ImageNet
- Se puede usar en entrenamiento profundidad estocástica: Antes de cada batch cierto numero de capas al azar se desactivan y se hacen equivalentes a la identidad



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112			$7 \times 7, 64, \text{stride } 2$		
conv2_x	56×56			$3 \times 3 \text{ max pool, stride } 2$		
conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1			average pool, 1000-d fc, softmax		
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

method	top-1 err.	top-5 err.
VGG [40] (ILSVRC'14)	-	8.43 [†]
GoogLeNet [43] (ILSVRC'14)	-	7.89
VGG [40] (v5)	24.4	7.1
PReLU-net [12]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

Top-5 error



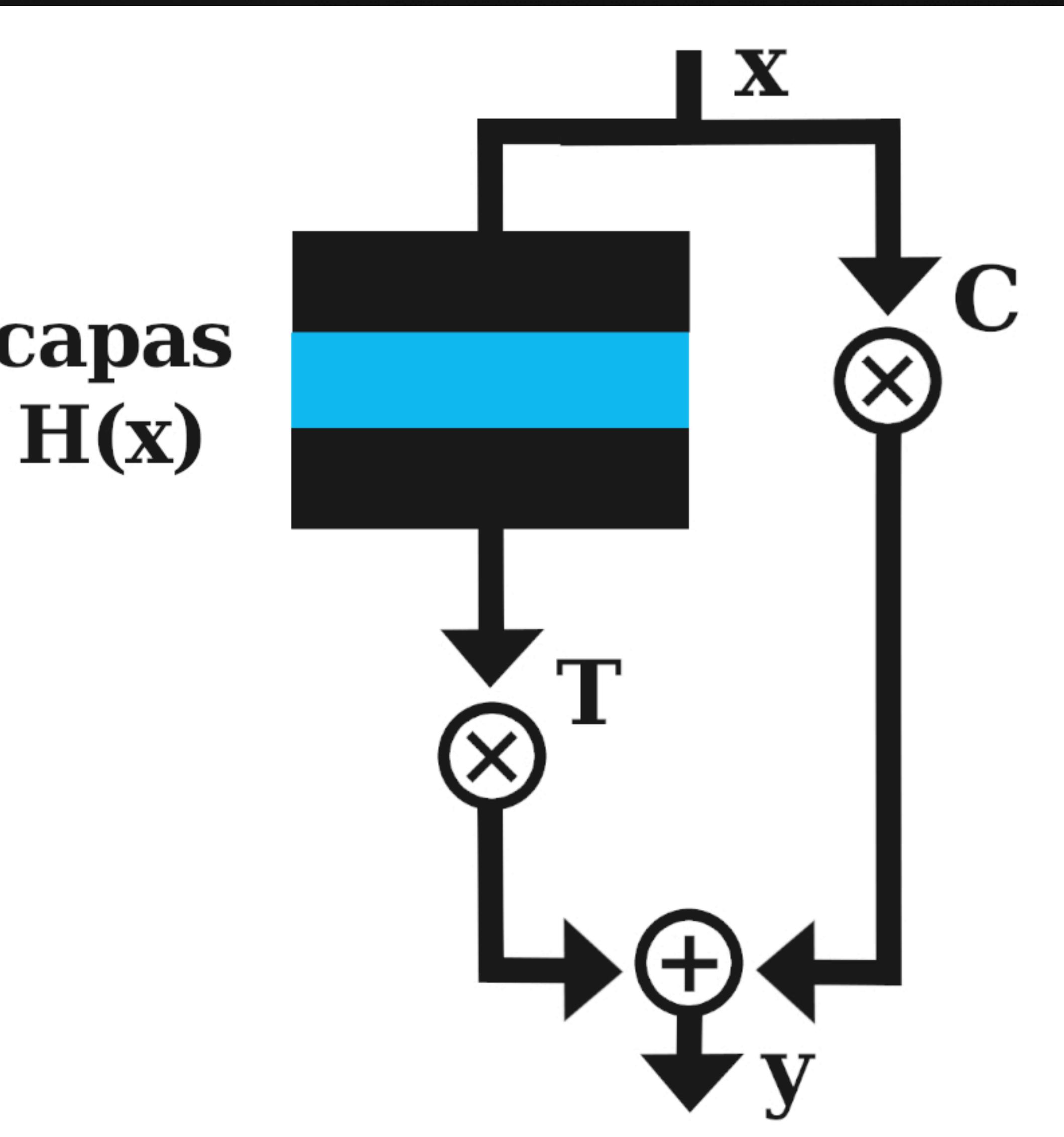
THAT'S NOT ENOUGH

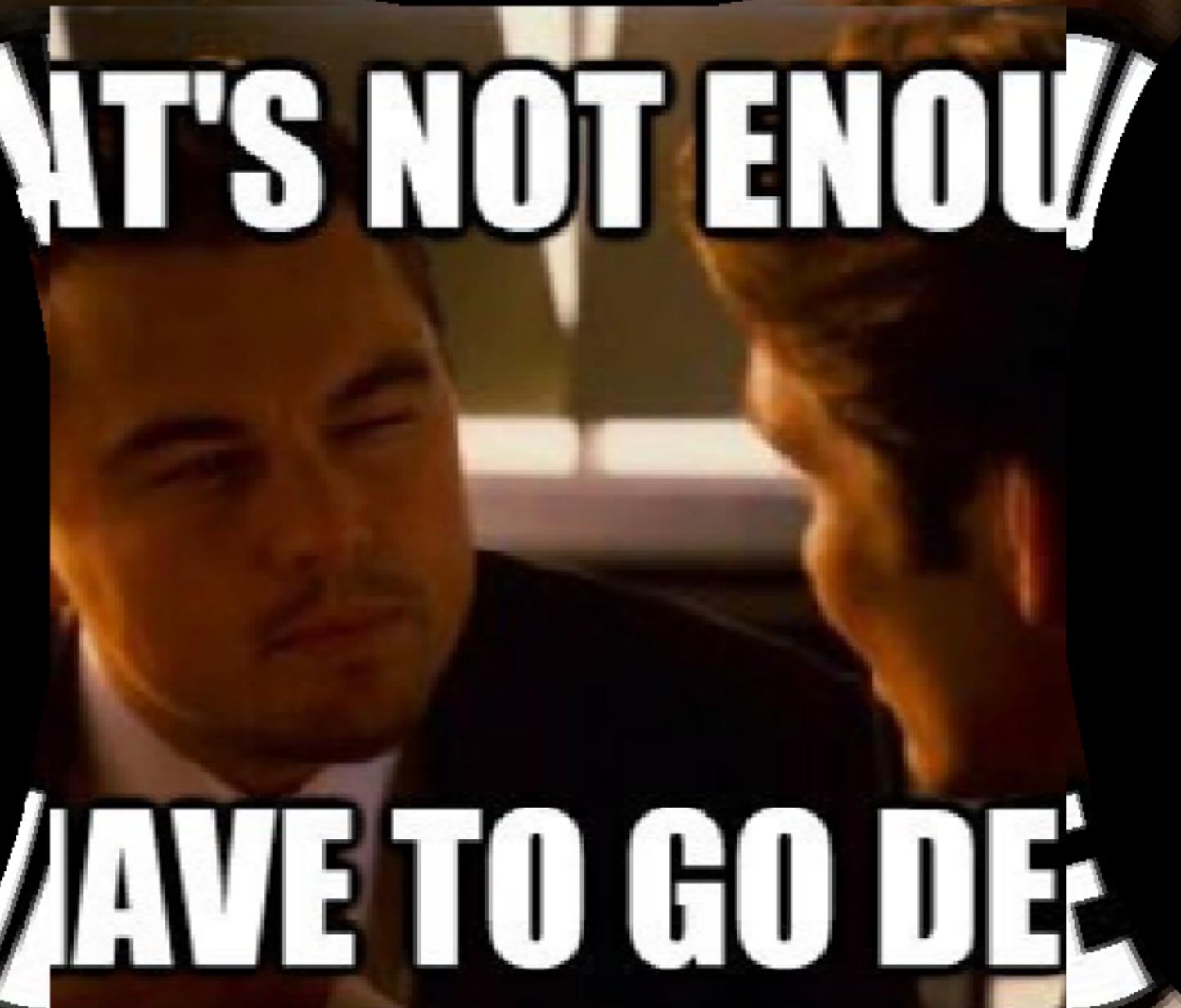
WE HAVE TO GO DEEPER

HighwayNet



- Se introducen dos compuertas tipo neurona la de transmisión T y la de acarreo C que parametrizan la cantidad de información que puede viajar a través de cada ruta
- La salida es $y_i = H_i(x)T_i(x) + xC_i(x)$ en donde $H_i(x)$ es equivalente al bloque residual de una ResNet
- Análisis en CIFAR-10 Mejoras de un orden de magnitud a 10 capas y posibilidad de ir hasta 100 capas sin degradarse





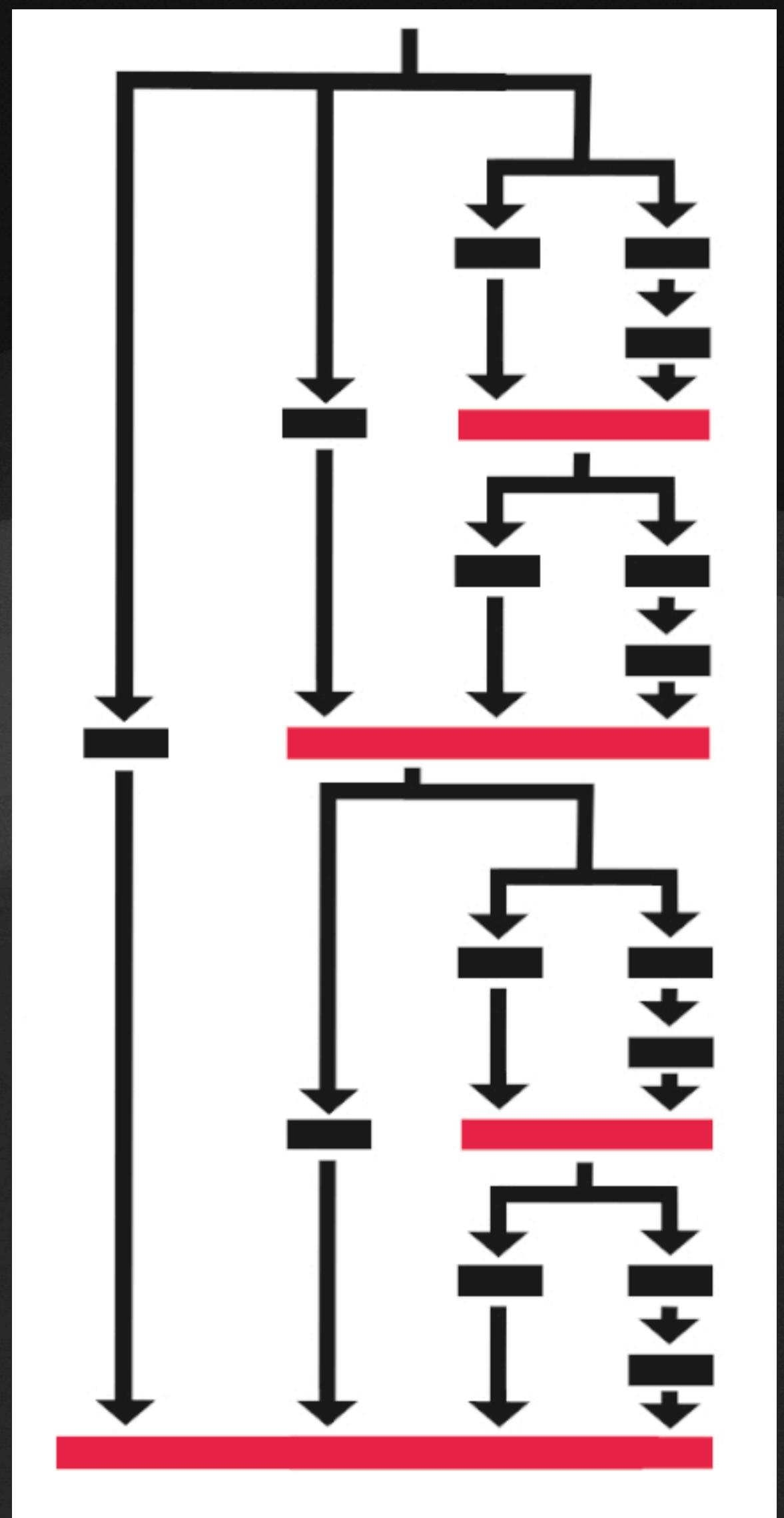
IT'S NOT ENOUGH

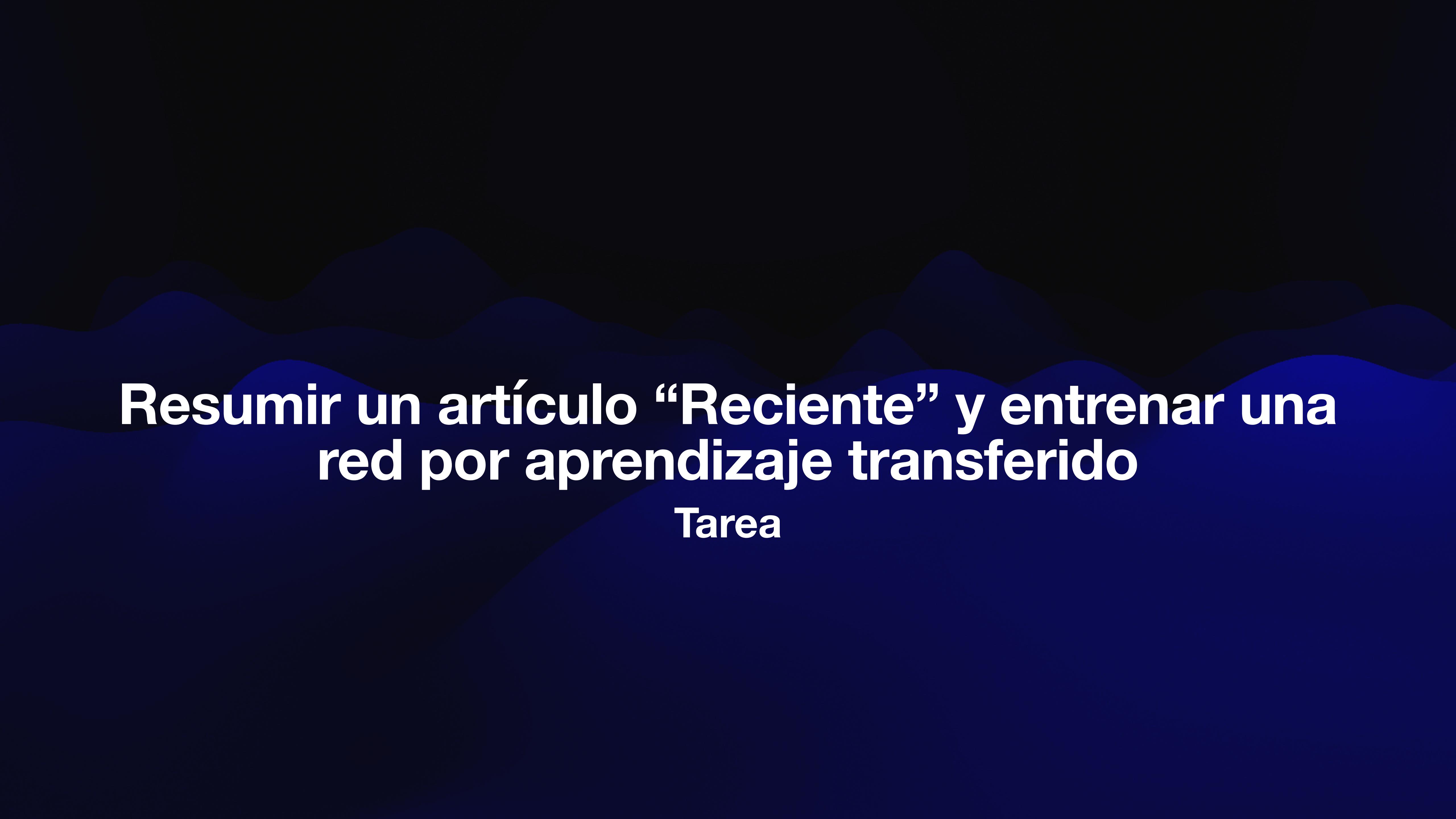
AVE TO GO DEE

FractalNet

<https://arxiv.org/pdf/1605.07648.pdf>

- Propuesta en 2017 por G Larsson, M. Maire y G. Shakhnarovich
- Bloques ensamblados autosemejantes
- Antes de entrenar cada batch se usa apagado local (aleatorio) o global (una sola ruta sobrevive)





Resumir un artículo “Reciente” y entrenar una
red por aprendizaje transferido

Tarea