Quién soy yo

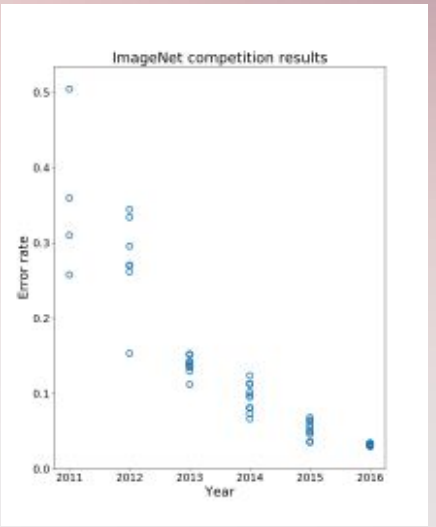
- Senior Data Scientist en Cognizant
- Especializado en Deep Learning y Computer Vision
- LinkedIn
- GitHub

Índice de la sesión

- Overview
- ResNet
- Vision Transformer
- Swin Transformer
- ResNeXt
- FocalNet
- Práctica

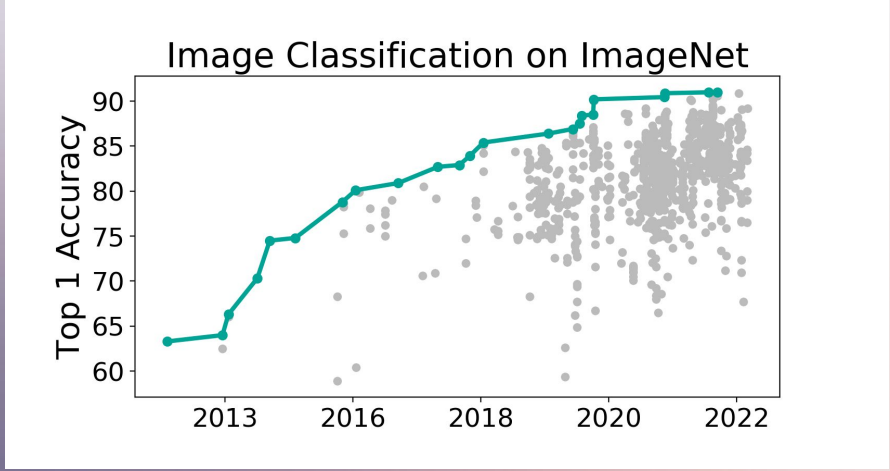
1. Overview

ImageNet es una dataset visual con más de 14 millones de imágenes anotadas manualmente.
Hay etiquetadas más de 20.000 categorías, información sobre los objetos en esa imagen y dónde están.



2. Overview

Transformers han ido ganando terreno en visión.
CNNs y ViTs se disputan el SOTA.

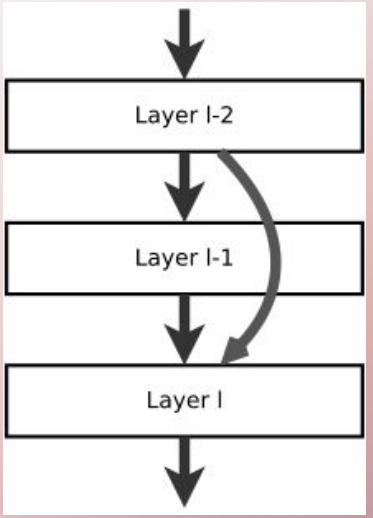


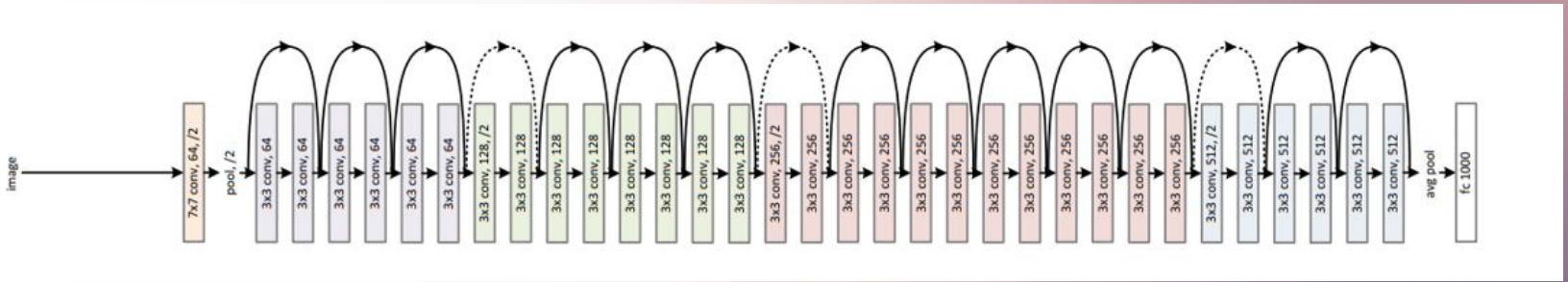
Model	#Params. (M)	FLOPs (G)	Throughput (imgs/s)	Top-1 (%)
ResNet-50 [30]	25.0	4.1	1294	76.2
ResNet-101 [30]	45.0	7.9	745	77.4
ResNet-152 [30]	60.0	11.0	522	78.3
ResNet-50-SB [88]	25.0	4.1	1294	79.8
ResNet-101-SB [88]	45.0	7.9	745	81.3
ResNet-152-SB [88]	60.0	11.6	522	81.8
DW-Net-T [28]	24.2	3.8	1030	81.2
DW-Net-B [28]	74.3	12.9	370	83.2
Mixer-B/16 [73]	59.9	12.7	455	76.4
gMLP-S [50]	19.5	4.5	785	79.6
gMLP-B [50]	73.4	15.8	301	81.6
ResMLP-S24 [74]	30.0	6.0	871	79.4
ResMLP-B24 [74]	129.1	23.0	61	81.0
DeiT-Small/16 [75]	22.1	4.6	939	79.9
DeiT-Base/16 [75]	86.6	17.5	291	81.8
PVT-Small [82]	24.5	3.8	794	79.8
PVT-Medium [82]	44.2	6.7	517	81.2
PVT-Large [82]	61.4	9.8	352	81.7
PoolFormer-m36 [100]	56.2	8.8	463	82.1
PoolFormer-m48 [100]	73.5	11.6	347	82.5
Swin-Tiny [54]	28.3	4.5	760	81.2
FocalNet-T (SRF)	28.4	4.4	743	**82.1**
Swin-Small [54]	49.6	8.7	435	83.1
FocalNet-S (SRF)	49.9	8.6	434	**83.4**
Swin-Base [54]	87.8	15.4	291	83.5
FocalNet-B (SRF)	88.1	15.3	280	**83.7**
FocalAtt-Tiny [95]	28.9	4.9	319	82.2
FocalNet-T (LRF)	28.6	4.5	696	**82.3**
FocalAtt-Small	51.1	9.4	192	**83.5**
FocalNet-S (LRF)	50.3	8.7	406	**83.5**
FocalAtt-Base [95]	89.8	16.4	138	83.8
FocalNet-B (LRF)	88.7	15.4	269	**83.9**

Table 1: ImageNet-1K classification comparison.

3. ResNet

Conexiones Residuales se utilizan para saltar sobre algunas capas.
Estas conexiones previenen desvaneci-miento de gradiente y hacen la red más fácil de entrenar.





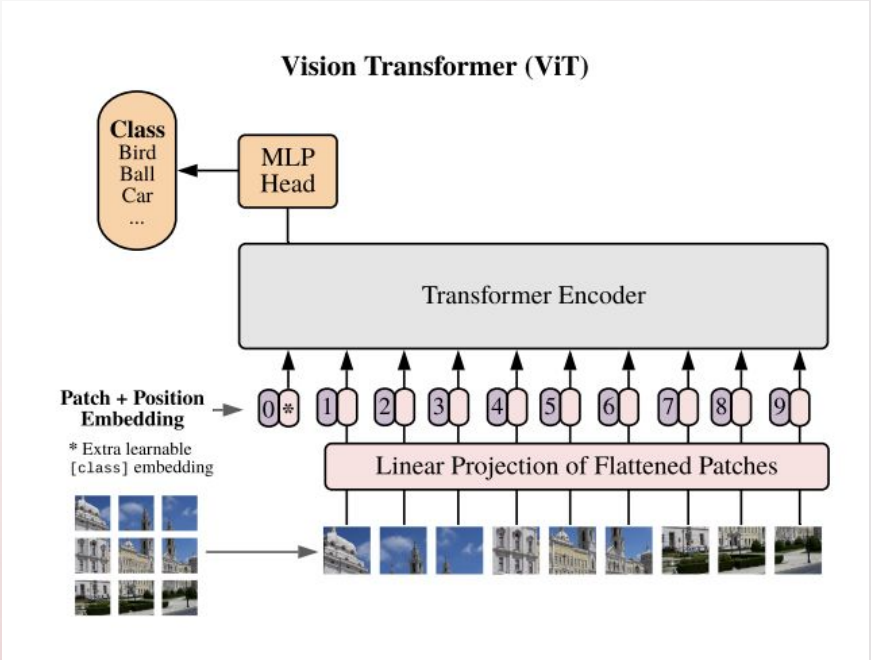
4. ResNet

Model	#Params. (M)	FLOPs (G)	Throughput (imgs/s)	Top-1 (%)
ResNet-50 [30]	25.0	4.1	1294	76.2
ResNet-101 [30]	45.0	7.9	745	77.4
ResNet-152 [30]	60.0	11.0	522	78.3
ResNet-50-SB [88]	25.0	4.1	1294	79.8
ResNet-101-SB [88]	45.0	7.9	745	81.3
ResNet-152-SB [88]	60.0	11.6	522	81.8
DW-Net-T [28]	24.2	3.8	1030	81.2
DW-Net-B [28]	74.3	12.9	370	83.2
Mixer-B/16 [73]	59.9	12.7	455	76.4
gMLP-S [50]	19.5	4.5	785	79.6
gMLP-B [50]	73.4	15.8	301	81.6
ResMLP-S24 [74]	30.0	6.0	871	79.4
ResMLP-B24 [74]	129.1	23.0	61	81.0
DeiT-Small/16 [75]	22.1	4.6	939	79.9
DeiT-Base/16 [75]	86.6	17.5	291	81.8
PVT-Small [82]	24.5	3.8	794	79.8
PVT-Medium [82]	44.2	6.7	517	81.2
PVT-Large [82]	61.4	9.8	352	81.7
PoolFormer-m36 [100]	56.2	8.8	463	82.1
PoolFormer-m48 [100]	73.5	11.6	347	82.5
Swin-Tiny [54]	28.3	4.5	760	81.2
FocalNet-T (SRF)	28.4	4.4	743	**82.1**
Swin-Small [54]	49.6	8.7	435	83.1
FocalNet-S (SRF)	49.9	8.6	434	**83.4**
Swin-Base [54]	87.8	15.4	291	83.5
FocalNet-B (SRF)	88.1	15.3	280	**83.7**
FocalAtt-Tiny [95]	28.9	4.9	319	82.2
FocalNet-T (LRF)	28.6	4.5	696	**82.3**
FocalAtt-Small	51.1	9.4	192	**83.5**
FocalNet-S (LRF)	50.3	8.7	406	**83.5**
FocalAtt-Base [95]	89.8	16.4	138	83.8
FocalNet-B (LRF)	88.7	15.4	269	**83.9**

Table 1: ImageNet-1K classification comparison.

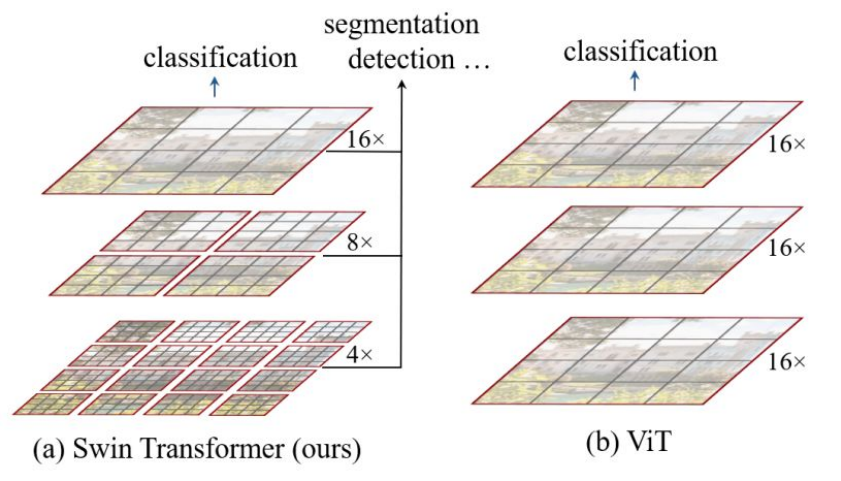
5. Vision Transformer

Los Transformers fueron creados inicialmente para texto.
Como en NLP se trata cada frase como una secuencia de palabras (o tokens), en una imagen se trata como una secuencia de parches.



6. Swin Transformer

Construye mapas de características jerárquicos mediante la fusión de parches de imágenes (que se muestran en gris) en capas más profundas.



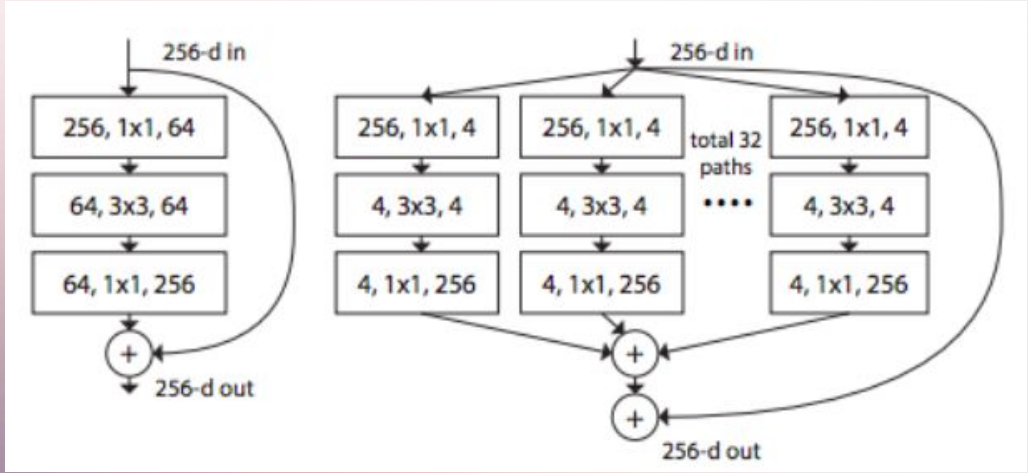
7. Swin Transformer

Model	#Params. (M)	FLOPs (G)	Throughput (imgs/s)	Top-1 (%)
ResNet-50 [30]	25.0	4.1	1294	76.2
ResNet-101 [30]	45.0	7.9	745	77.4
ResNet-152 [30]	60.0	11.0	522	78.3
ResNet-50-SB [88]	25.0	4.1	1294	79.8
ResNet-101-SB [88]	45.0	7.9	745	81.3
ResNet-152-SB [88]	60.0	11.6	522	81.8
DW-Net-T [28]	24.2	3.8	1030	81.2
DW-Net-B [28]	74.3	12.9	370	83.2
Mixer-B/16 [73]	59.9	12.7	455	76.4
gMLP-S [50]	19.5	4.5	785	79.6
gMLP-B [50]	73.4	15.8	301	81.6
ResMLP-S24 [74]	30.0	6.0	871	79.4
ResMLP-B24 [74]	129.1	23.0	61	81.0
DeiT-Small/16 [75]	22.1	4.6	939	79.9
DeiT-Base/16 [75]	86.6	17.5	291	81.8
PVT-Small [82]	24.5	3.8	794	79.8
PVT-Medium [82]	44.2	6.7	517	81.2
PVT-Large [82]	61.4	9.8	352	81.7
PoolFormer-m36 [100]	56.2	8.8	463	82.1
PoolFormer-m48 [100]	73.5	11.6	347	82.5
Swin-Tiny [54]	28.3	4.5	760	81.2
FocalNet-T (SRF)	28.4	4.4	743	**82.1**
Swin-Small [54]	49.6	8.7	435	83.1
FocalNet-S (SRF)	49.9	8.6	434	**83.4**
Swin-Base [54]	87.8	15.4	291	83.5
FocalNet-B (SRF)	88.1	15.3	280	**83.7**
FocalAtt-Tiny [95]	28.9	4.9	319	82.2
FocalNet-T (LRF)	28.6	4.5	696	**82.3**
FocalAtt-Small	51.1	9.4	192	**83.5**
FocalNet-S (LRF)	50.3	8.7	406	**83.5**
FocalAtt-Base [95]	89.8	16.4	138	83.8
FocalNet-B (LRF)	88.7	15.4	269	**83.9**

Table 1: ImageNet-1K classification comparison.

8. ResNeXt

La nueva generación de ResNets. Más eficientes y potentes.
Bloques con misma topología se unen.

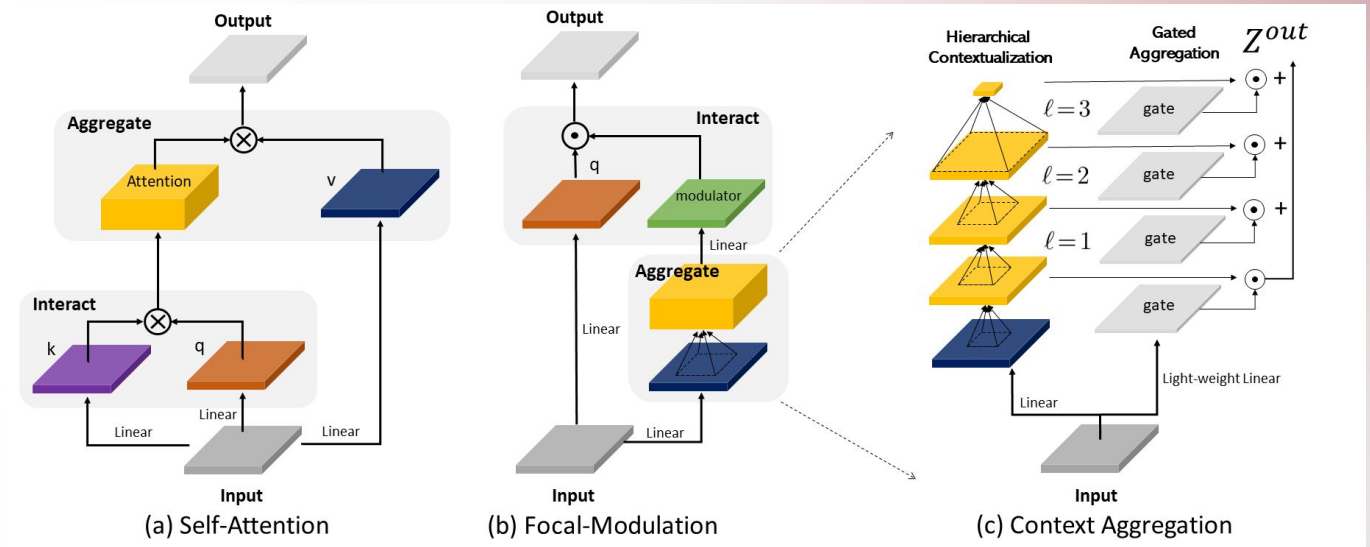


9. ResNeXt

Models	Params	FLOPS	Acc.
ResNet-164 (Baseline)	1.74M	0.25G	77.0
ResNet-200	2.11M	0.30G	77.5
ResNet-164 (CPWC w/o Stage 2)	1.87M	0.28G	**77.7**
ResNet-164 CPWC (Ours)	1.96M	0.30G	**78.4**
ResNeXt-29,8×64d (Baseline)	34.4M	5.4G	82.23
ResNeXt-29,16×64d	68.1M	10.7G	82.69
ResNeXt-29,8×64d CPWC (Ours)	34.8M	5.5G	**82.64**
MobileNetV2(1.0x) (Baseline)	2.4M	0.30G	74.1
MobileNetV2(1.2x)	3.2M	0.41G	74.3
MobileNetV2(1.0x) CPWC (Ours)	2.6M	0.35G	**75.1**

10. FocalNet

Usa Convoluciones con diferentes kernels para crear diferentes niveles de representación de imágenes que luego son agregadas de forma adaptativa. Esto sustituye al mecanismo de atención de los Transformers.



11. FocalNet

Model	#Params. (M)	FLOPs (G)	Throughput (imgs/s)	Top-1 (%)
ResNet-50 [30]	25.0	4.1	1294	76.2
ResNet-101 [30]	45.0	7.9	745	77.4
ResNet-152 [30]	60.0	11.0	522	78.3
ResNet-50-SB [88]	25.0	4.1	1294	79.8
ResNet-101-SB [88]	45.0	7.9	745	81.3
ResNet-152-SB [88]	60.0	11.6	522	81.8
DW-Net-T [28]	24.2	3.8	1030	81.2
DW-Net-B [28]	74.3	12.9	370	83.2
Mixer-B/16 [73]	59.9	12.7	455	76.4
gMLP-S [50]	19.5	4.5	785	79.6
gMLP-B [50]	73.4	15.8	301	81.6
ResMLP-S24 [74]	30.0	6.0	871	79.4
ResMLP-B24 [74]	129.1	23.0	61	81.0
DeiT-Small/16 [75]	22.1	4.6	939	79.9
DeiT-Base/16 [75]	86.6	17.5	291	81.8
PVT-Small [82]	24.5	3.8	794	79.8
PVT-Medium [82]	44.2	6.7	517	81.2
PVT-Large [82]	61.4	9.8	352	81.7
PoolFormer-m36 [100]	56.2	8.8	463	82.1
PoolFormer-m48 [100]	73.5	11.6	347	82.5
Swin-Tiny [54]	28.3	4.5	760	81.2
FocalNet-T (SRF)	28.4	4.4	743	**82.1**
Swin-Small [54]	49.6	8.7	435	83.1
FocalNet-S (SRF)	49.9	8.6	434	**83.4**
Swin-Base [54]	87.8	15.4	291	83.5
FocalNet-B (SRF)	88.1	15.3	280	**83.7**
FocalAtt-Tiny [95]	28.9	4.9	319	82.2
FocalNet-T (LRF)	28.6	4.5	696	**82.3**
FocalAtt-Small	51.1	9.4	192	**83.5**
FocalNet-S (LRF)	50.3	8.7	406	**83.5**
FocalAtt-Base [95]	89.8	16.4	138	83.8
FocalNet-B (LRF)	88.7	15.4	269	**83.9**

Table 1: ImageNet-1K classification comparison.

12. Práctica

Vamos a usar la librería TorchVision para usar los modelos anteriores preentrenados con ImageNet 1k.

LINK

pillow

colab

¡Muchas gracias por uniros a esta W3 Drop!