Redes Neuronales

Temas avazandos

Agenda

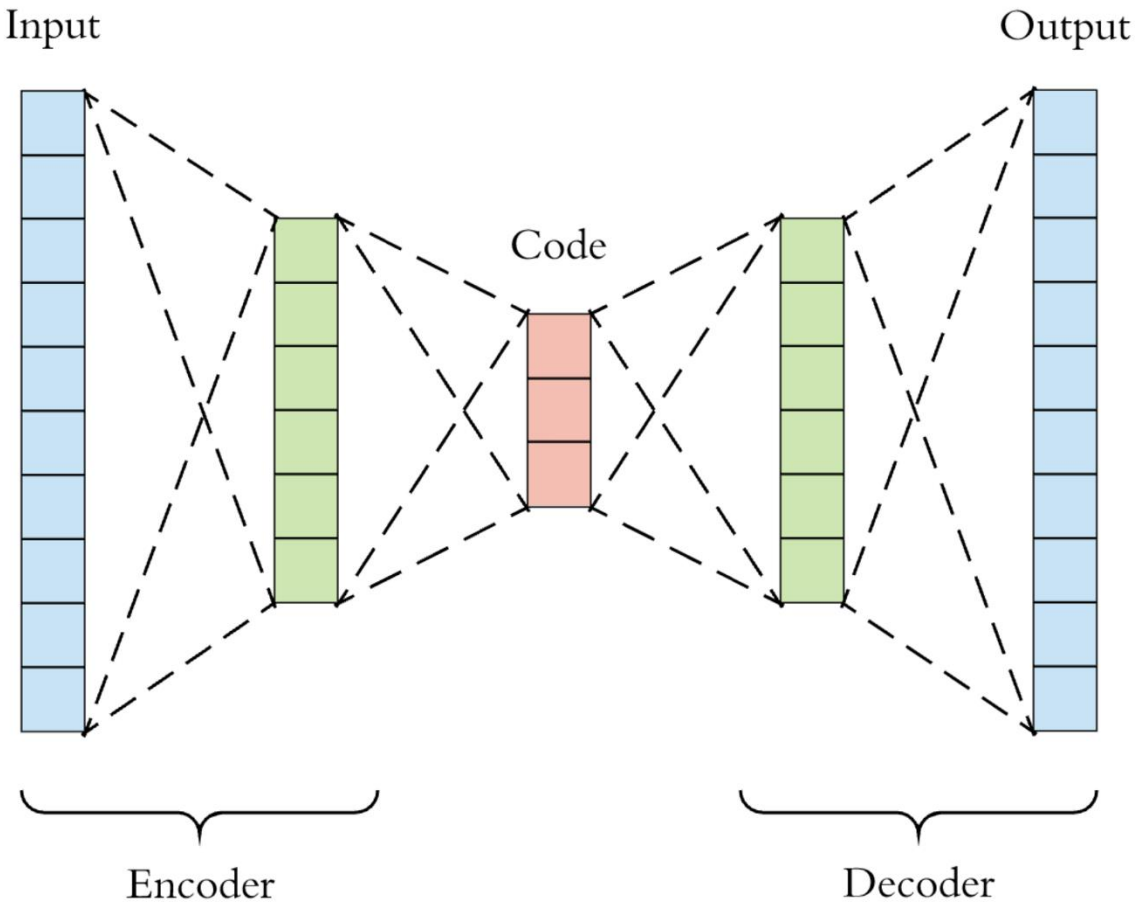
- Autoencoder
- Generative Adversarials Networks
- Transfer Learning
- Quantization
- Sparsification
- Conclusiones

Autoencoders

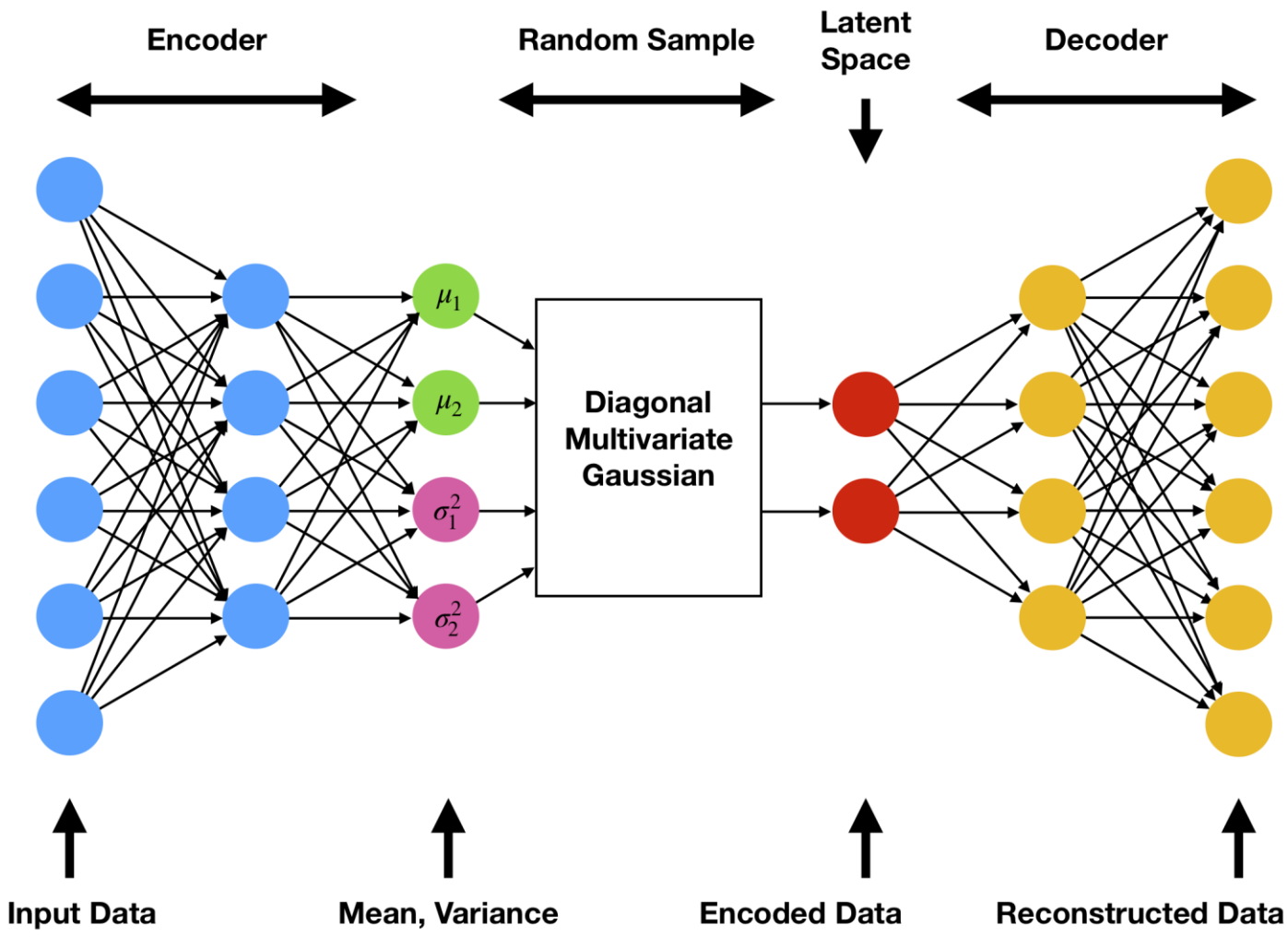
Autoencoders son útiles para:

- Reducción de dimensionalidad
- Reducción de ruido
- Generación automática
- Recomendación

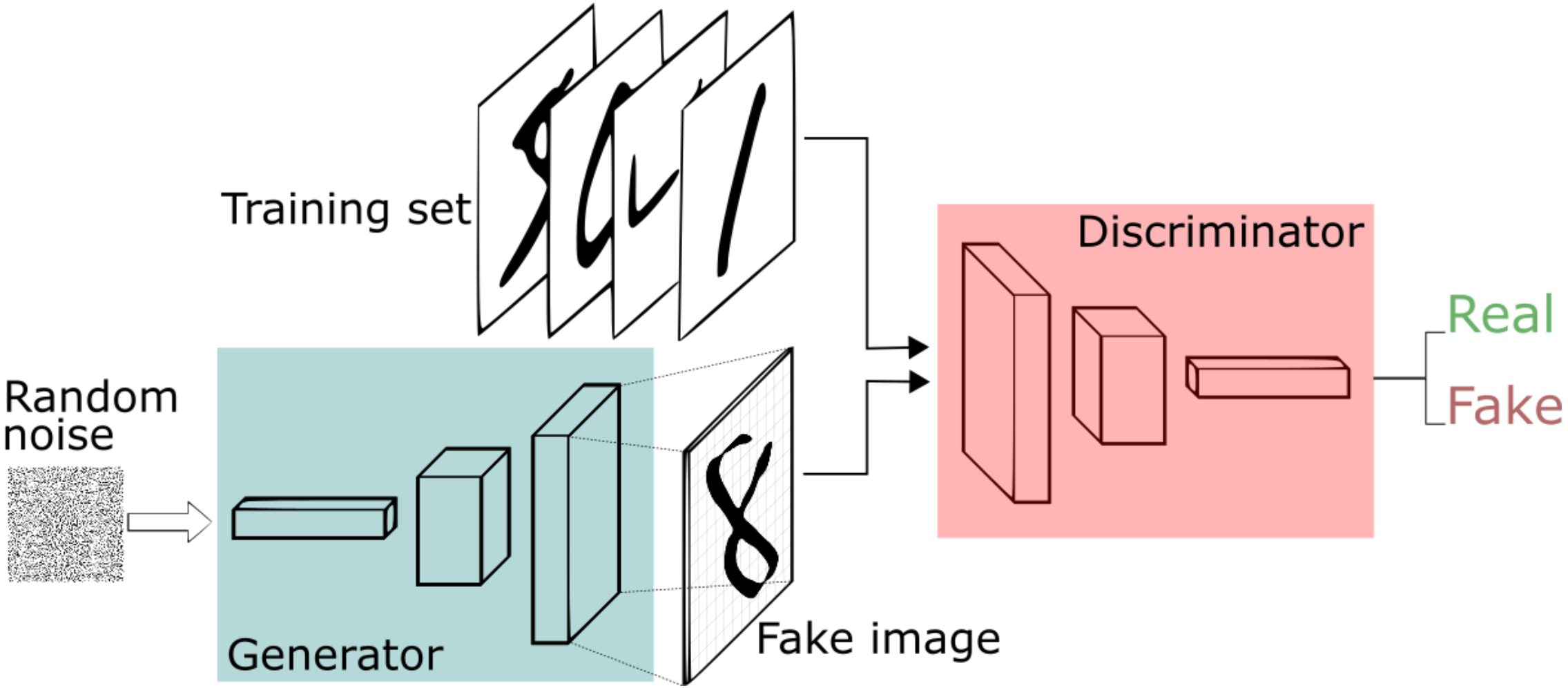
Autoencoders



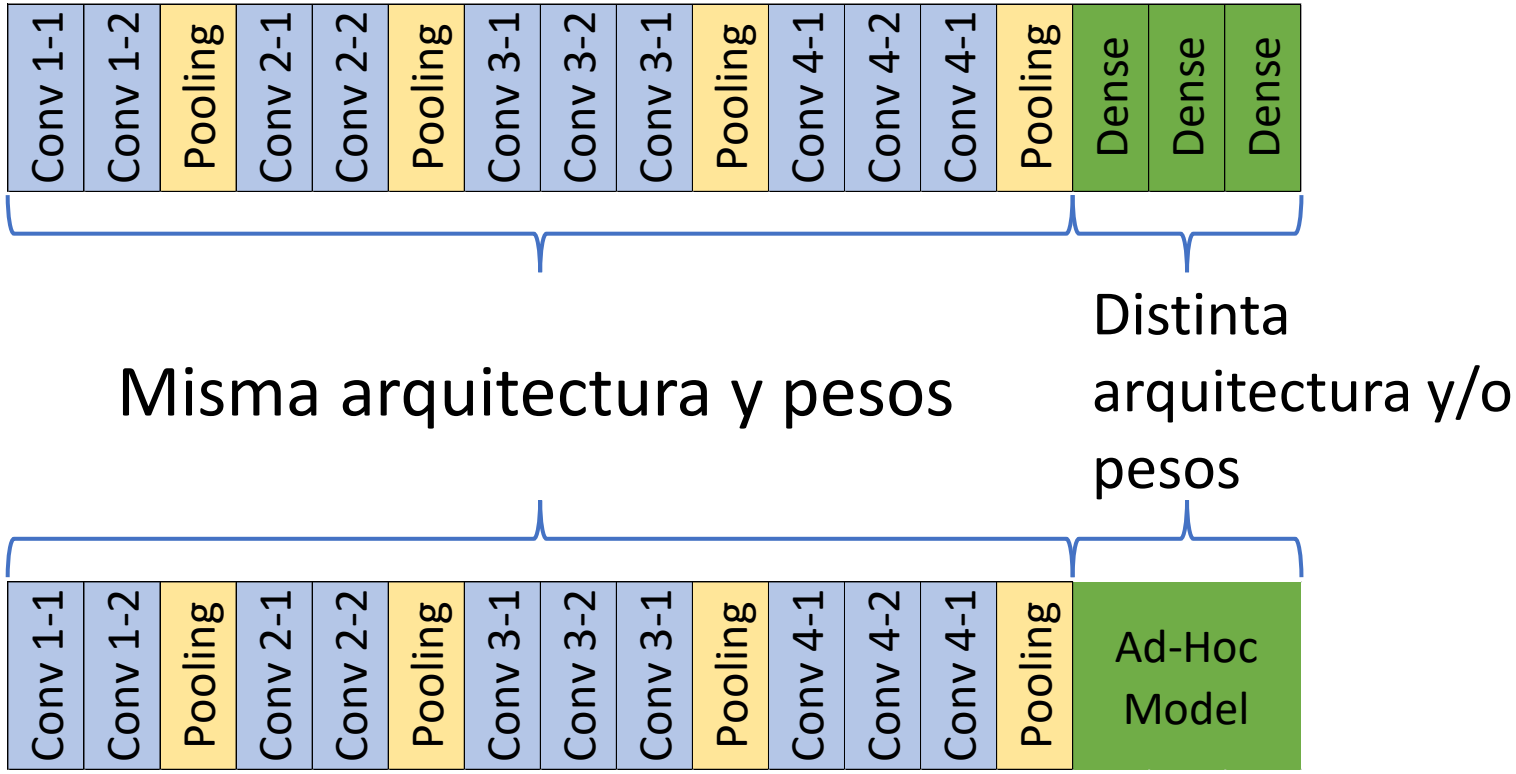
Autoencoders



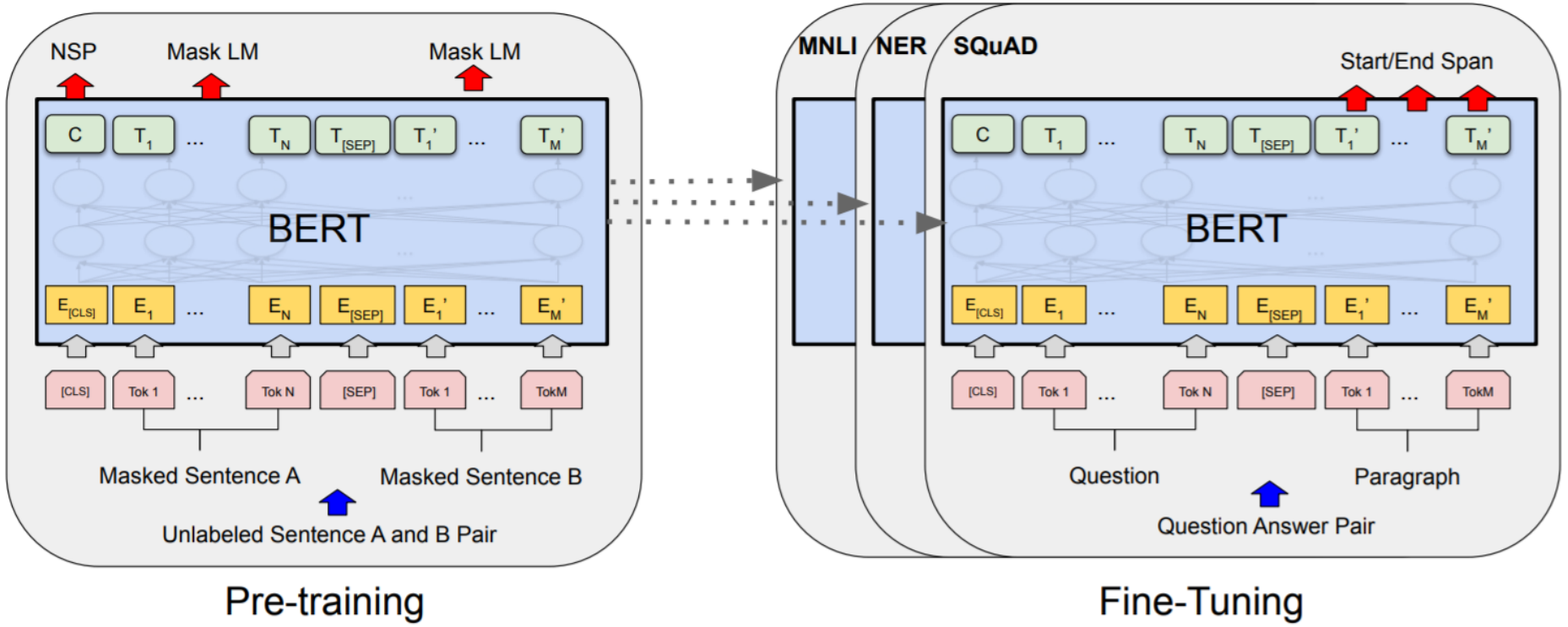
GANs


Training set
Discriminator
Real
Fake
Random noise
Generator
Fake image

Transfer learning

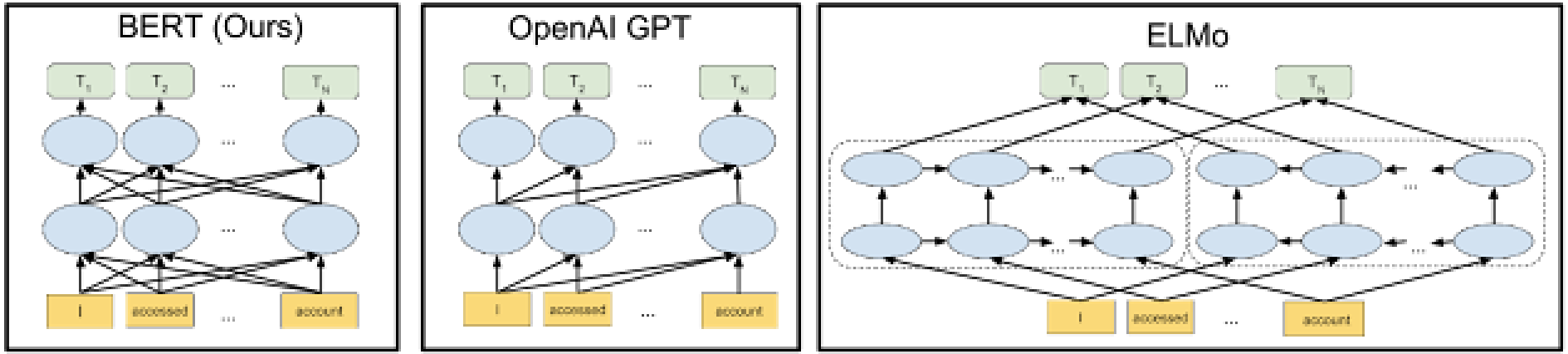


Transfer learning: BERT



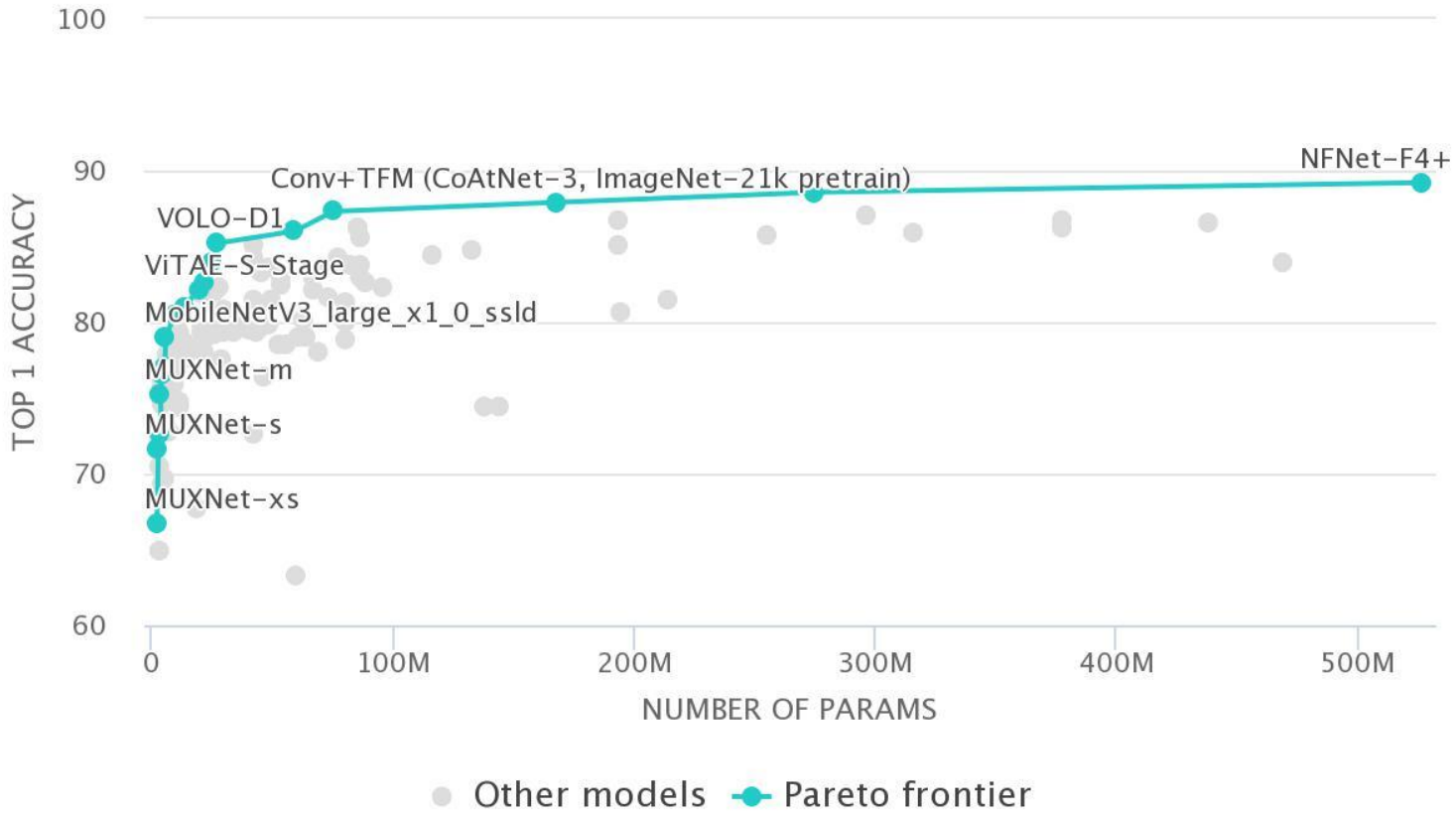
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
https://arxiv.org/abs/1810.04805

Transfer learning: ¡más modelos!



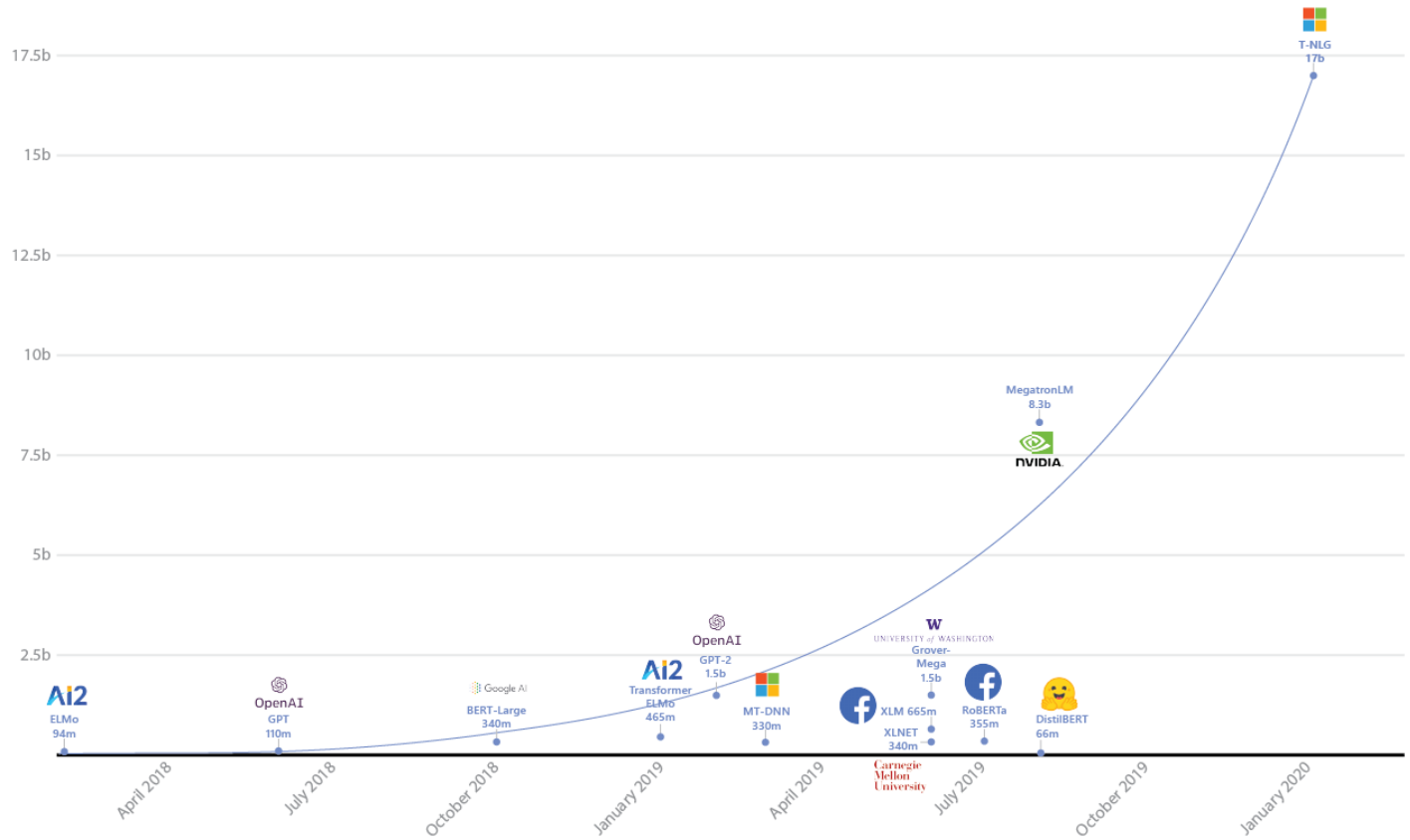
https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Transfer learning: + datos + parámetros!



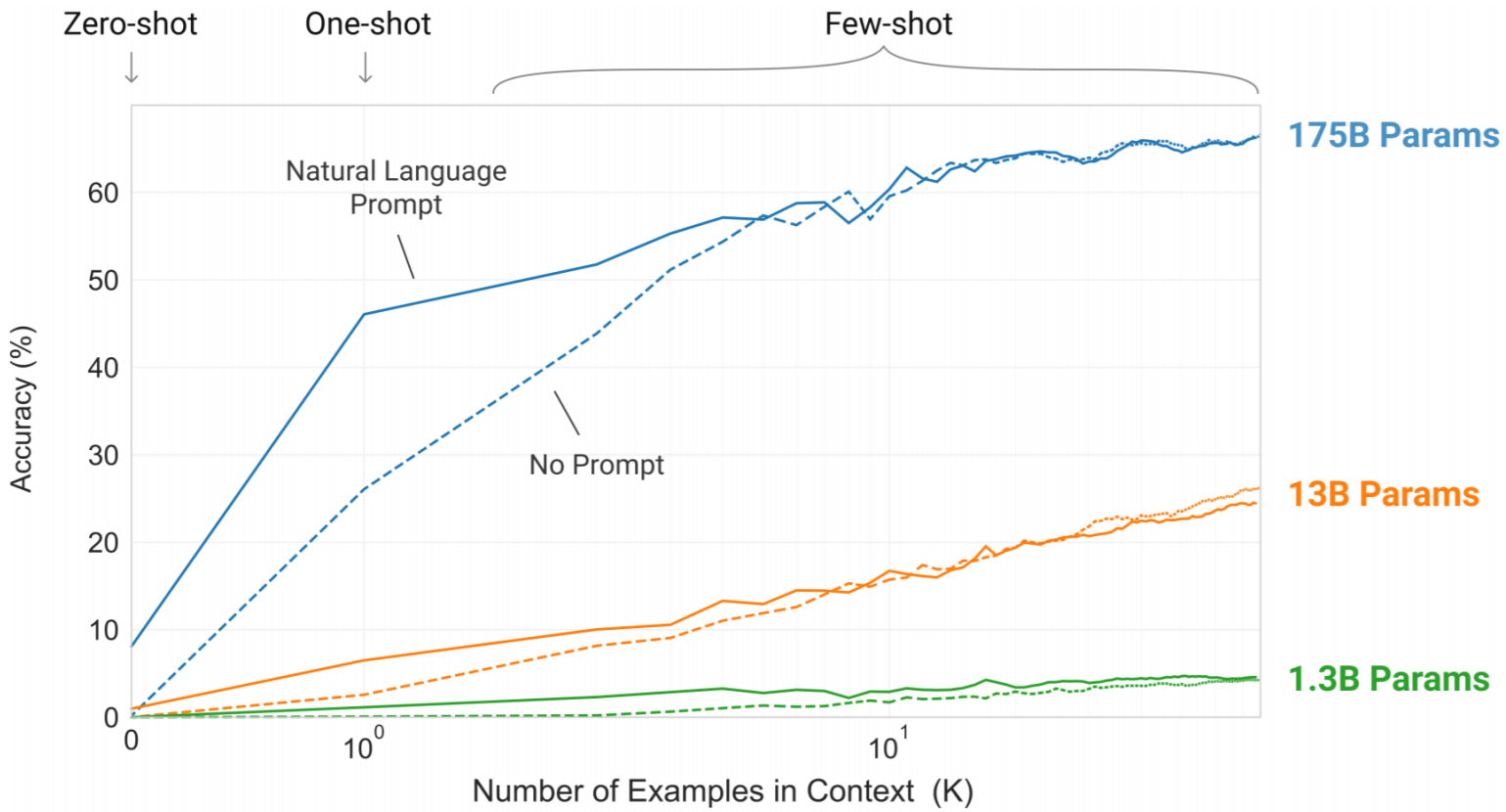
https://paperswithcode.com/sota/image-classification-on-imagenet?dimension=Number%20of%20params&tag_filter=0

Transfer learning: + datos + parámetros!



https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Transfer learning: + datos + parámetros!



Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165..

Quantization

Technique	Data requirements	Size reduction	Accuracy	Supported hardware
Post-training float16 quantization	No data	Up to 50%	Insignificant accuracy loss	CPU, GPU
Post-training dynamic range quantization	No data	Up to 75%	Accuracy loss	CPU, GPU (Android)
Post-training integer quantization	Unlabelled representative sample	Up to 75%	Smaller accuracy loss	CPU, GPU (Android), EdgeTPU, Hexagon DSP
Quantization-aware training	Labelled training data	Up to 75%	Smallest accuracy loss	CPU, GPU (Android), EdgeTPU, Hexagon DSP

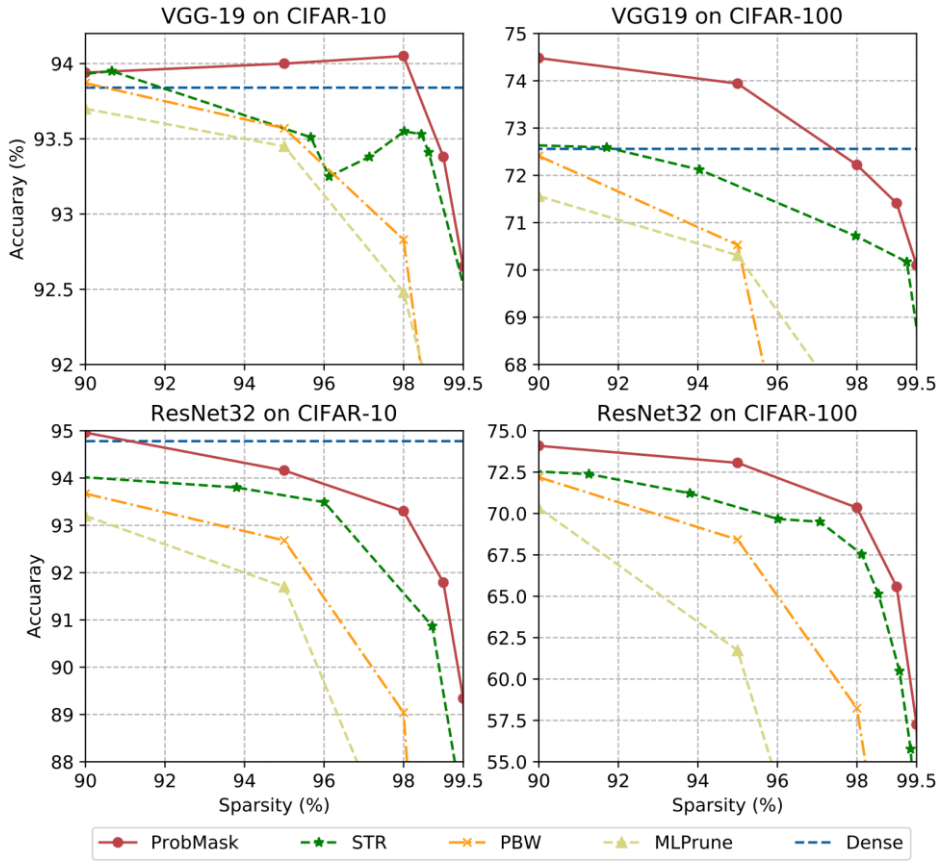
https://www.tensorflow.org/lite/performance/model_optimization#quantization

Quantization

Model	Top-1 Accuracy (Original)	Top-1 Accuracy (Post Training Quantized)	Top-1 Accuracy (Quantization Aware Training)	Latency (Original) (ms)	Latency (Post Training Quantized) (ms)	Latency (Quantization Aware Training) (ms)	Size (Original) (MB)	Size (Optimized) (MB)
Mobilenet-v1-1-224	0.709	0.657	0.70	124	112	64	16.9	4.3
Mobilenet-v2-1-224	0.719	0.637	0.709	89	98	54	14	3.6
Inception_v3	0.78	0.772	0.775	1130	845	543	95.7	23.9
Resnet_v2_101	0.770	0.768	N/A	3973	2868	N/A	178.3	44.9

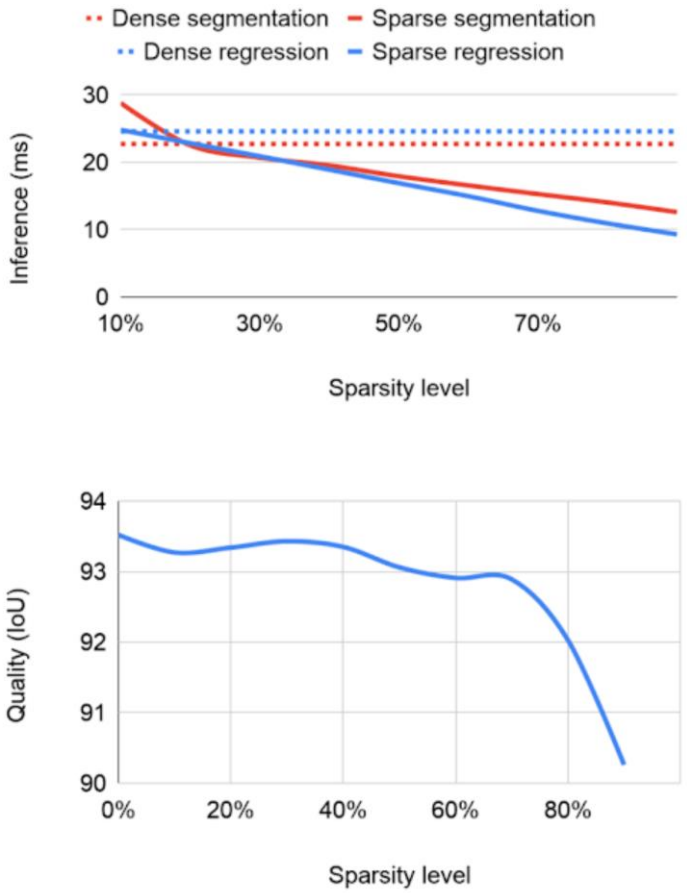
https://www.tensorflow.org/lite/performance/model_optimization#quantization

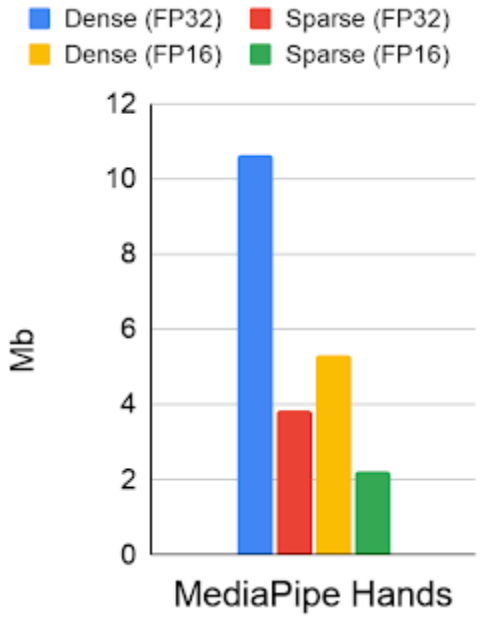
Spasification

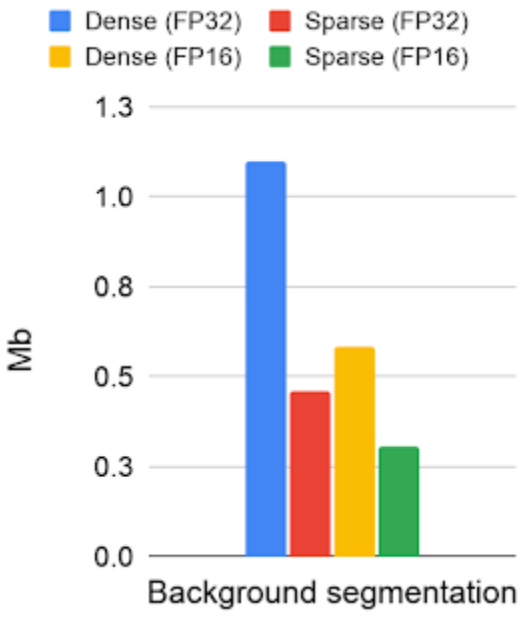


https://openaccess.thecvf.com/content/CVPR2021/papers/Zhou_Effective_Sparsification_of_Neural_Networks_With_Global_Sparsity_Constraint_CVPR_2021_paper.pdf

Spasification







https://ai.googleblog.com/2021/03/accelerating-neural-networks-on-mobile.html

Herramientas y más...

- Tensorflow
- Pytorch
- Apache mxnet
- ONNX
- HuggingFace
- NVIDIA TensorRT
- NVIDIA Triton Inference Server
- Tensorflow Serving
- Tensorflow Lite
- …