



Self-supervised Learning with Transformers networks focused on Computer Vision tasks

Cristhian Wiki Sánchez Sauñé

csanchezs@uni.pe

Facultad de Ciencias

07 de Enero, 2022



Tabla de Contenido

1. Generalidades

Problemática

Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)

Self Supervised Learning (SSL)

Destilación

Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento

Clasificación k -NN

Mapas *self-attention*

Segmentación

Demo interactiva

5. Conclusiones y Trabajos Futuros

Tabla de Contenido

1. Generalidades

Problemática
Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)
Self Supervised Learning (SSL)
Destilación
Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento
Clasificación k -NN
Mapas *self-attention*
Segmentación
Demo interactiva

5. Conclusiones y Trabajos Futuros



Generalidades

Problemática

- Redes Convolucionales no es todo lo que necesitas: Vision Transformers [Dosovitskiy et al., 2020]





Generalidades

Problemática

- Redes Convolucionales no es todo lo que necesitas: Vision Transformers [Dosovitskiy et al., 2020]
- Redes Neuronales como cajas negras





Generalidades

Problemática

- Redes Convolucionales no es todo lo que necesitas: Vision Transformers [Dosovitskiy et al., 2020]
- Redes Neuronales como cajas negras
- Los datos en la vida real se encuentran NO etiquetados





Generalidades

Objetivo General

Evaluar el desempeño del algoritmo DINO frente a otros algoritmos *self-supervised* actuales para diversas tareas de Computer Vision.



Generalidades

Objetivos Específicos

- O1) Analizar la explicabilidad e interpretabilidad de DINO frente a otros métodos.
- O2) Realizar un estudio de ablación del algoritmo DINO, analizando las consecuencias de modificar diversos hiperparámetros tales como el *batchsize*, función de pérdida, etc.

Tabla de Contenido

1. Generalidades

Problemática
Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)
Self Supervised Learning (SSL)
Destilación
Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento
Clasificación k -NN
Mapas *self-attention*
Segmentación
Demo interactiva

5. Conclusiones y Trabajos Futuros

Marco Conceptual

Vision Transformer (ViT)

Modelo para la clasificación de imágenes que emplea una arquitectura similar al Transformer original [Vaswani et al., 2017] sobre parches de una imagen.

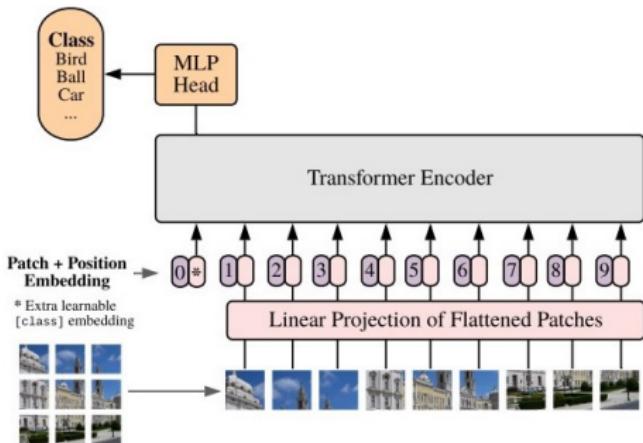


Figure: Arquitectura ViT presentada en [Dosovitskiy et al., 2020]

Marco Conceptual Self Supervised Learning (SSL)

Es una técnica de aprendizaje relativamente reciente que a diferencia del aprendizaje supervisado, no requiere ningún dato etiquetado. En su lugar, crea pseudoetiquetas autodefinidas como señales de supervisión y aprende representaciones, que luego se utilizan en tareas posteriores.

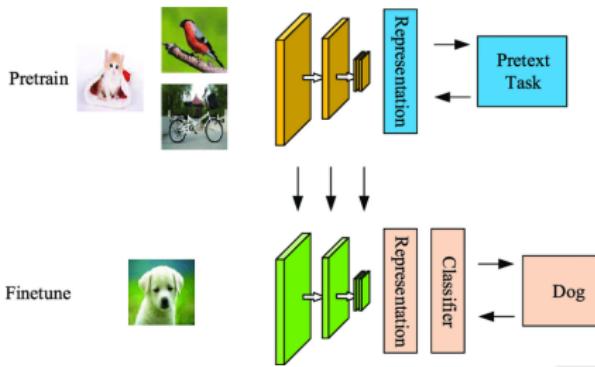


Figure: Flujo usual de diversos métodos *self-supervised*.

Marco Conceptual

Destilación de modelos

Es un procedimiento para la compresión de un modelo, en la que se entrena un modelo pequeño (estudiante) para que intente replicar la salida de un gran modelo pre-entrenado (maestro).

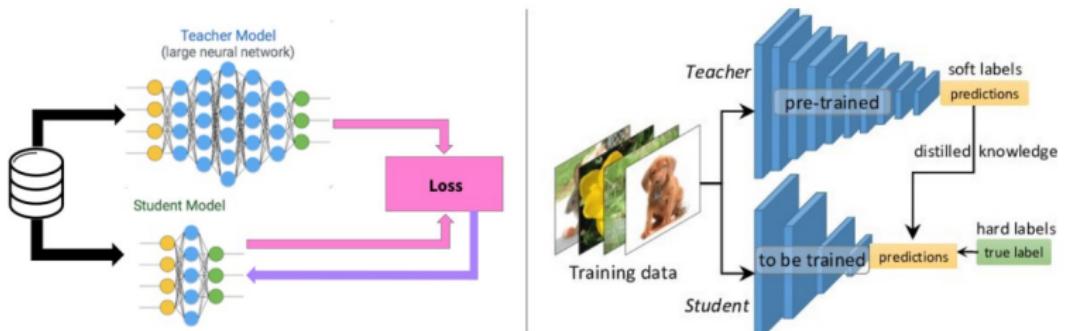


Figure: Flujo completo del proceso de destilación de un modelo

Marco Conceptual

Media Móvil Exponencial (EMA)

$$V_t = \beta * V_{t-1} + (1 - \beta) * \theta_t$$

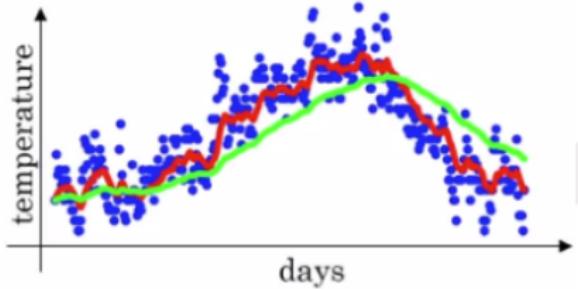


Figure: Visualización de la media para diferentes valores de β .

Estado del Arte

Análisis de rendimiento

A continuación se muestra un cuadro comparativo de las diferentes arquitecturas SOA, extraído de [Li et al., 2021]:

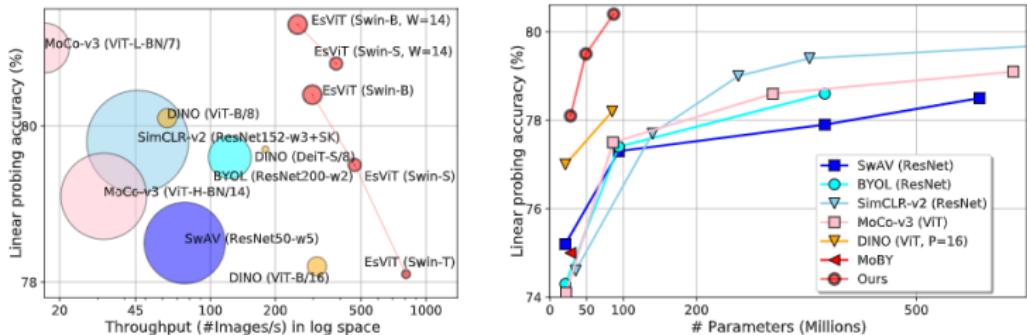


Figure: Comparación del rendimiento vs accuracy bajo el protocolo de clasificación lineal en ImageNet.

Tabla de Contenido

1. Generalidades

Problemática
Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)
Self Supervised Learning (SSL)
Destilación
Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento
Clasificación k -NN
Mapas *self-attention*
Segmentación
Demo interactiva

5. Conclusiones y Trabajos Futuros

Metodología



Student

Teacher

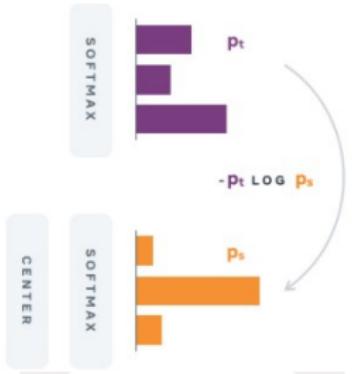


Figure: Paso forward usando el algoritmo DINO.

Metodología

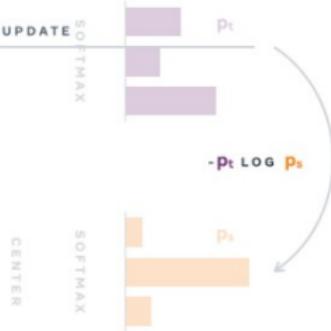
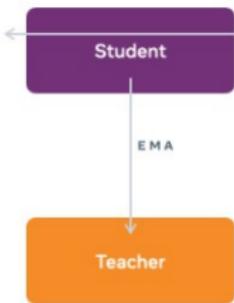


Figure: Paso backward usando el algoritmo DINO.

Metodología

SSL con destilación del conocimiento

La probabilidad P se obtiene normalizando la salida de la red con una función softmax.

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^K \exp(g_{\theta_s}(x)^{(k)}/\tau_s)} \quad (1)$$

donde $\tau_s > 0$ es un parámetro de temperatura que controla la diversidad de la salida, y una fórmula similar se mantiene para P_t con temperatura τ_t .

Metodología

SSL con destilación del conocimiento

Las redes aprenderán mediante la minimización de la pérdida de entropía cruzada H :

$$\min_{\theta_s} H(P_t(x), P_s(x)) \quad (2)$$

donde $H(a, b) = -a \log b$.

La parte de centrado se usa para evitar que el maestro pueda diverger:

$$C \leftarrow mC + (1 - m) \frac{1}{B} \sum_{i=1}^B g_{\theta_t}(x_i) \quad (3)$$

donde B es el *batchsize*.

Metodología

Arquitectura de red

Se usaron las redes DeiT (variante ViT) y ResNet por poseer muchas similitudes, tanto en cantidad de parámetros como en tiempo de entrenamiento.

Modelo	bloques	dim.	heads	#tokens	#parám
ResNet-50	-	2048	-	-	23M
DeiT-S/16	12	384	6	197	21M
DeiT-S/8	12	384	6	785	21M
ResNet-50	12	768	12	197	85M
ResNet-50	12	768	12	785	85M

Herramientas

Hardware y Software

- ① Para la creación de los modelos, el framework Pytorch.
- ② Para la visualización de diversas métricas durante el entrenamiento se usó Tensorboard.
- ③ Para el entrenamiento, una gráfica NVIDIA 1050Ti.



Fuente: Elaboración propia.

Herramientas

Datasets: ImageNet

Para las tareas más generales de visión se usó ImageNet [Russakovsky et al., 2015], que contiene 14197,122 imágenes anotadas con 1k etiquetas.

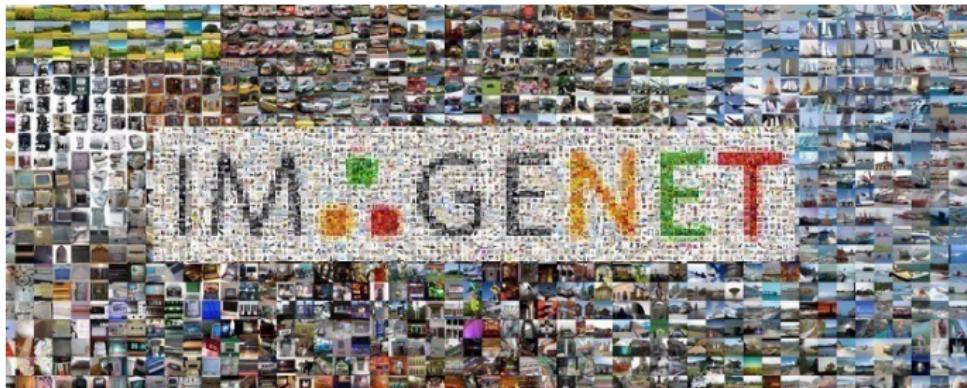


Figure: Imagen extraída de [Russakovsky et al., 2015].

Herramientas

Datasets: DAVIS 2017 Video object segmentation

Para la segmentación de objetos se usó el dataset DAVIS (Densely Annotated Video Segmentation) [Perazzi et al., 2016, Pont-Tuset et al., 2017], el cual consiste en secuencias de video Full HD de alta calidad. Cada video está acompañado por una segmentación anotada, precisa de píxeles y por cuadro.



Figure: Imagen extraída de [Perazzi et al., 2016].

Herramientas

Pascal VOC

Para obtener métricas de validación en imágenes de segmentación, aparte del dataset mencionado en el ítem anterior, también se usó el dataset PASCAL Visual Object Classes (VOC) 2012 [Everingham et al.].



Figure: Imagen extraída de [Everingham et al.].

Tabla de Contenido

1. Generalidades

Problemática
Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)
Self Supervised Learning (SSL)
Destilación
Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento

Clasificación k -NN

Mapas *self-attention*

Segmentación

Demo interactiva

5. Conclusiones y Trabajos Futuros

Experimentación

Visualización de métricas durante el entrenamiento

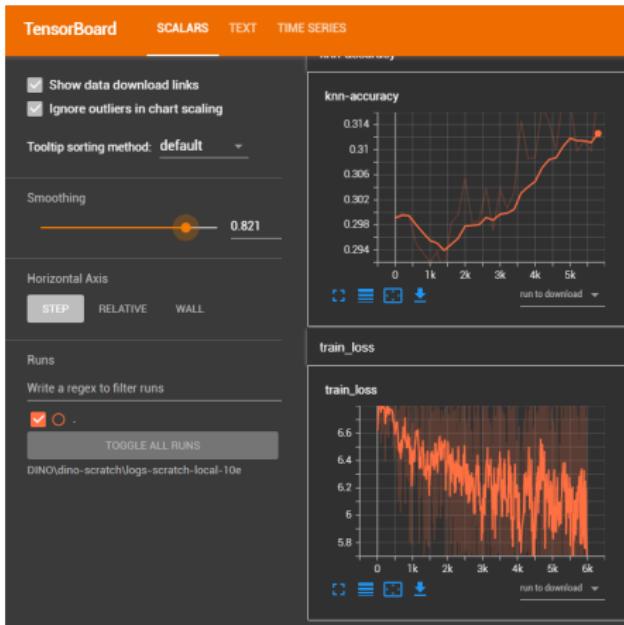


Figure: Métricas durante las primeras 8 épocas.

Resultados

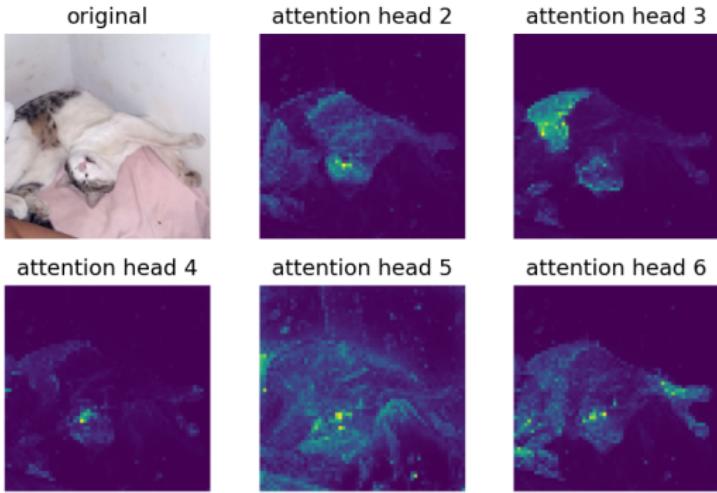
Clasificación k -NN en ImageNet

Método	Arquit.	Parám.	im/s	Lineal	k -NN
Supervisado	RN50	23	637	79.2	79.2
SCLR [Chen et al., 2020a]	RN50	23	637	69.0	60.6
MoCov2 [Chen et al., 2020b]	RN50	23	637	71.2	61.8
BYOL [Grill et al., 2020]	RN50	23	637	74.2	64.4
SwAV [Caron et al., 2020]	RN50	23	637	75.1	65.2
DINO	RN50	23	637	75.1	67.4
Supervisado	DeiT-S	21	503	79.5	79.5
BYOL [Grill et al., 2020]	DeiT-S	21	503	71.5	66.7
MoCov2 [Chen et al., 2020b]	DeiT-S	21	503	72.9	64.5
SwAV [Caron et al., 2020]	DeiT-S	21	503	73.6	66.5
DINO	DeiT-S	21	503	77.1	74.4

Resultados

Mapas *self-attention*

Estos mapas muestran que el modelo aprende automáticamente características específicas de la clase que conducen a segmentaciones de objetos de forma *self-supervised*.



Resultados

Segmentación de forma supervisada versus DINO

Se muestran las máscaras resultantes de un DeiT-S/8 entrenado de forma supervisada y con DINO.

Supervised



DINO



Resultados

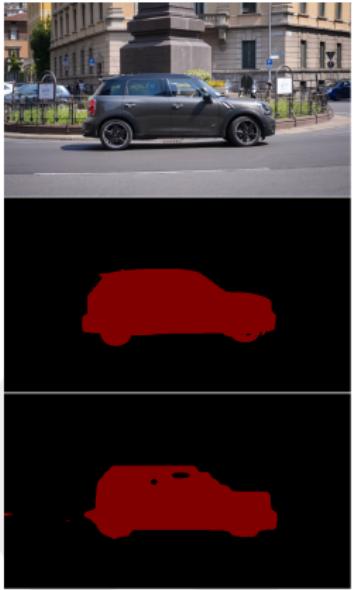
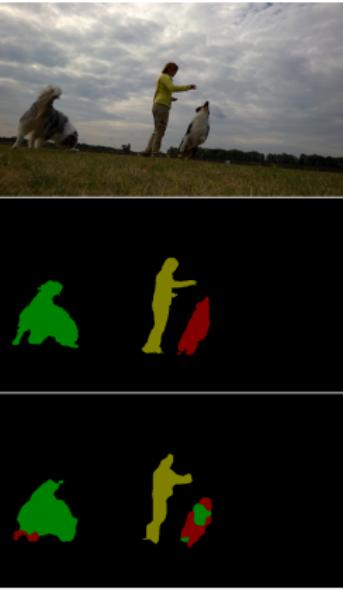
Segmentación de forma supervisada versus DINO

Comparación de la similitud Jaccard entre el ground truth y las máscaras generadas. Se usó el dataset PASCAL VOC [Everingham et al.].

	Random	Supervisado	DINO
DeiT-S/16	19.8	26.9	45.8
DeiT-S/8	20.7	23.6	44.6

Resultados

Segmentación de instancias en video usando DAVIS 2017



Resultados

Demo interactiva para poder visualizar los mapas de atención

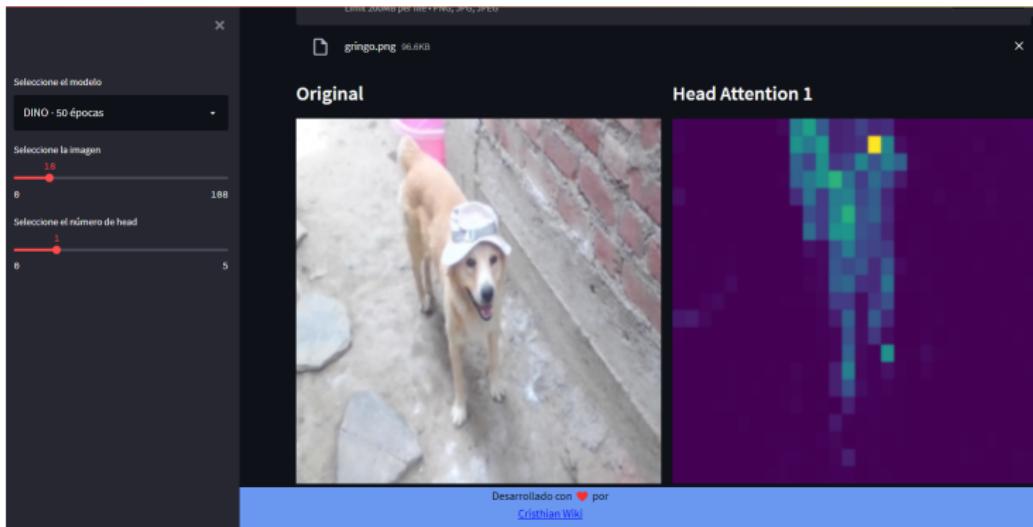


Figure: Aplicación de demostración.

Tabla de Contenido

1. Generalidades

Problemática

Objetivos

2. Marco Conceptual y SOA

Vision Transformer (ViT)

Self Supervised Learning (SSL)

Destilación

Media Móvil Exponencial (EMA)

Estado del Arte

3. Metodología y Herramientas

4. Experimentación y Resultados

Entrenamiento

Clasificación k -NN

Mapas *self-attention*

Segmentación

Demo interactiva

5. Conclusiones y Trabajos Futuros

Conclusiones

Se observa que emergen 3 propiedades que pueden aprovecharse en aplicaciones futuras:

- La calidad de las características son muy útiles para una clasificación k -NN.



Conclusiones

Se observa que emergen 3 propiedades que pueden aprovecharse en aplicaciones futuras:

- La calidad de las características son muy útiles para una clasificación k -NN.
- La presencia de información sobre el diseño de la escena también puede beneficiar a la segmentación de imágenes débilmente supervisada.



Conclusiones

Se observa que emergen 3 propiedades que pueden aprovecharse en aplicaciones futuras:

- La calidad de las características son muy útiles para una clasificación k -NN.
- La presencia de información sobre el diseño de la escena también puede beneficiar a la segmentación de imágenes débilmente supervisada.
- Hay una clara sinergia entre DINO y ViT, comparado con otros métodos *self-supervised* / *contrastive*.

Trabajo futuro

- Explorar si el pre-entrenamiento de un gran modelo ViT con DINO en imágenes aleatorias sin tratar podría empujar los límites de las características visuales [Goyal et al., 2021].



Trabajo futuro

- Explorar si el pre-entrenamiento de un gran modelo ViT con DINO en imágenes aleatorias sin tratar podría empujar los límites de las características visuales [Goyal et al., 2021].
- Explorar el enfoque self-supervised en imágenes de análisis médico, buscando producir un impacto más directo y tangible como resultado de usar estos novedosos métodos en casos reales.



Referencias bibliográficas I

- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.
- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.
<http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- Priya Goyal, Mathilde Caron, Benjamin Lefauzeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.



Referencias bibliográficas II

Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Muchas gracias!

