

Vision por Computador

Modelos Fundacionales

Agenda



1. Modelos fundamentales
2. Self-Supervised Learning
3. Discusión
4. Taller

Agenda



- 1. Modelos fundacionales**
2. Self-Supervised Learning
3. Discusión
4. Taller

Foundation Models

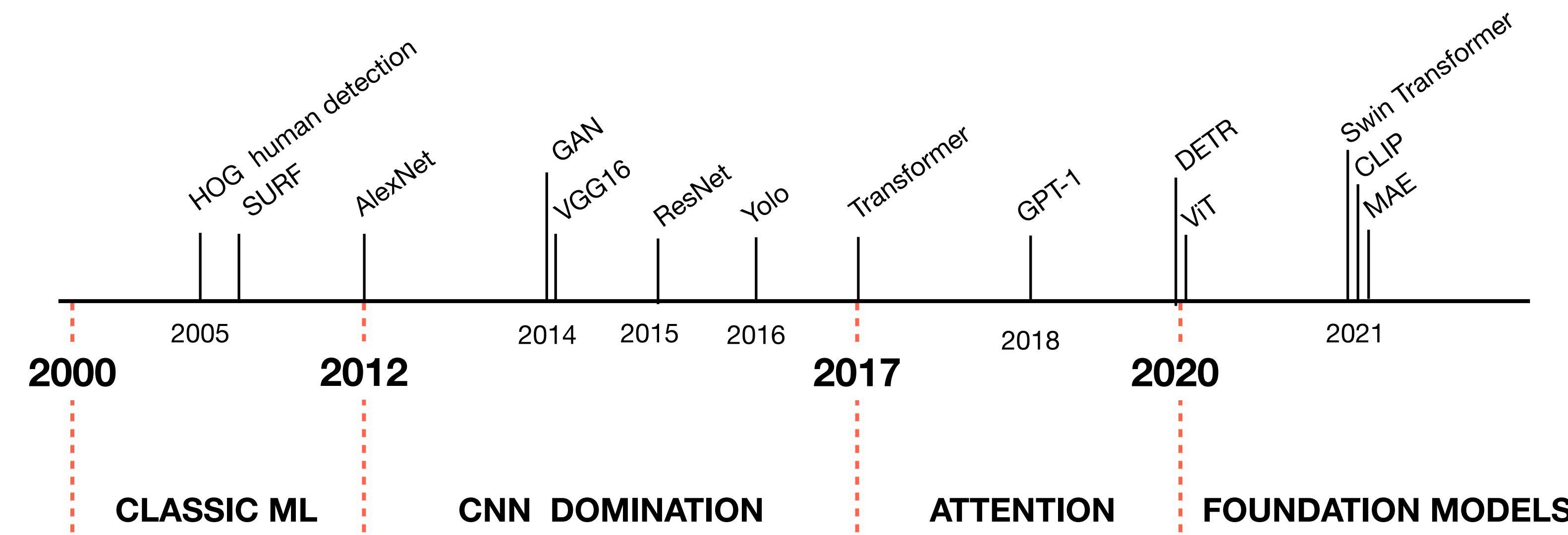
Modelos Fundacionales

1. Un modelo de ML o DL entrenados en grandes conjuntos de datos de **gran escala y diversidad**
2. Diseñados para desempeñarse en una gran **variedad de tareas y dominios**
3. Proporcionan una sólida inicialización de parámetros para una amplia gama de **aplicaciones posteriores** (Downstream)

Model Type	Dataset Size	Examples
Traditional CNNs	Millions of labeled images	ImageNet (1.2M), COCO (330K)
Foundational Models	Hundreds of millions to billions	CLIP (400M), GPT-4 Vision (unspecified, likely billions)

Línea de tiempo

Foundation Models



Elementos Clave

Foundation Models

- A. Conjuntos de datos masivos y diversos
- B. Agnóstico a tareas
- C. Arquitectura flexible
- D. Aprendizaje autosupervisado
- E. Centrado en la representación



Agenda

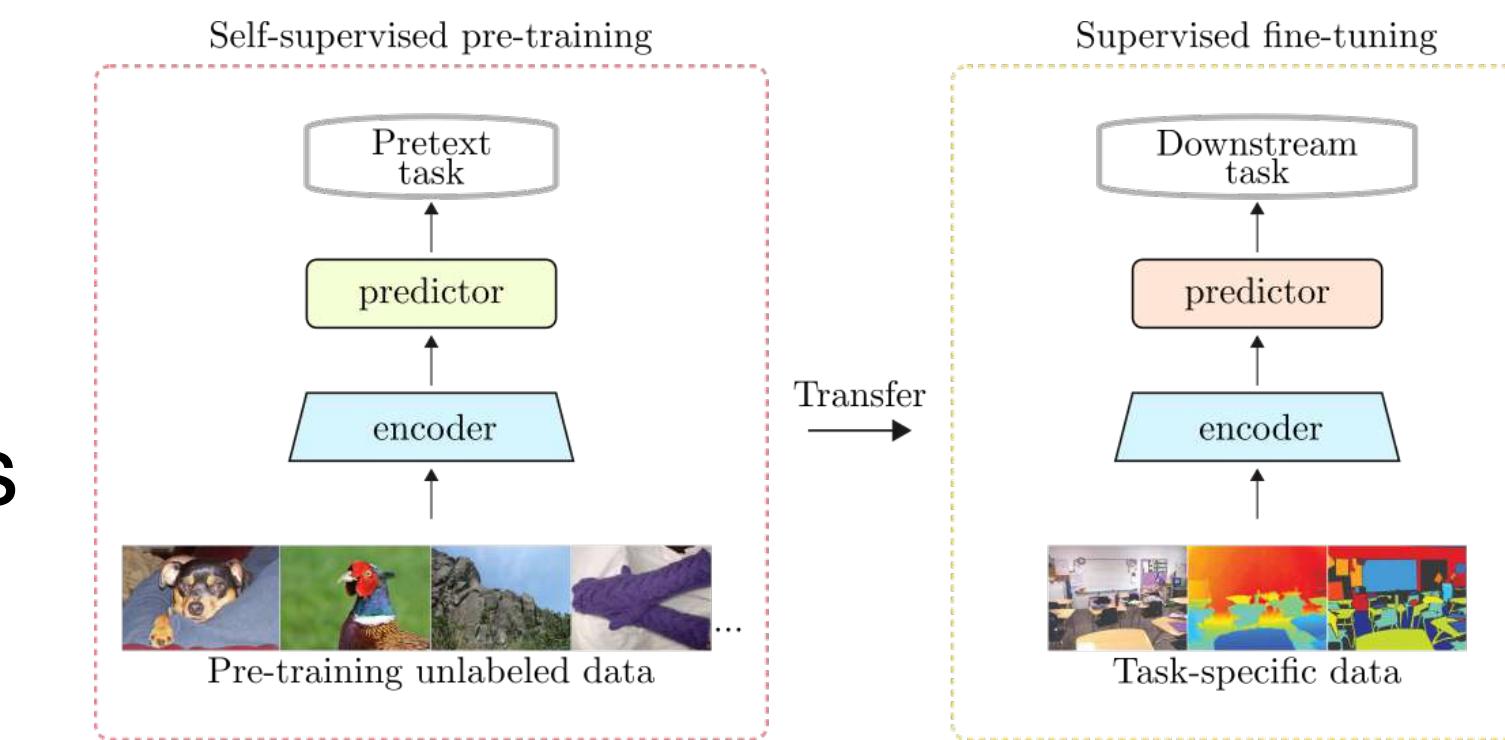
- Modelos fundamentales
- **Self-Supervised Learning**
 - Predictivos - Pretext tasks
 - Contrastivos (SimSLR)
 - Generativos (MAE)
- Discusión
- Taller



Principios básicos

Aprendizaje autosupervisado

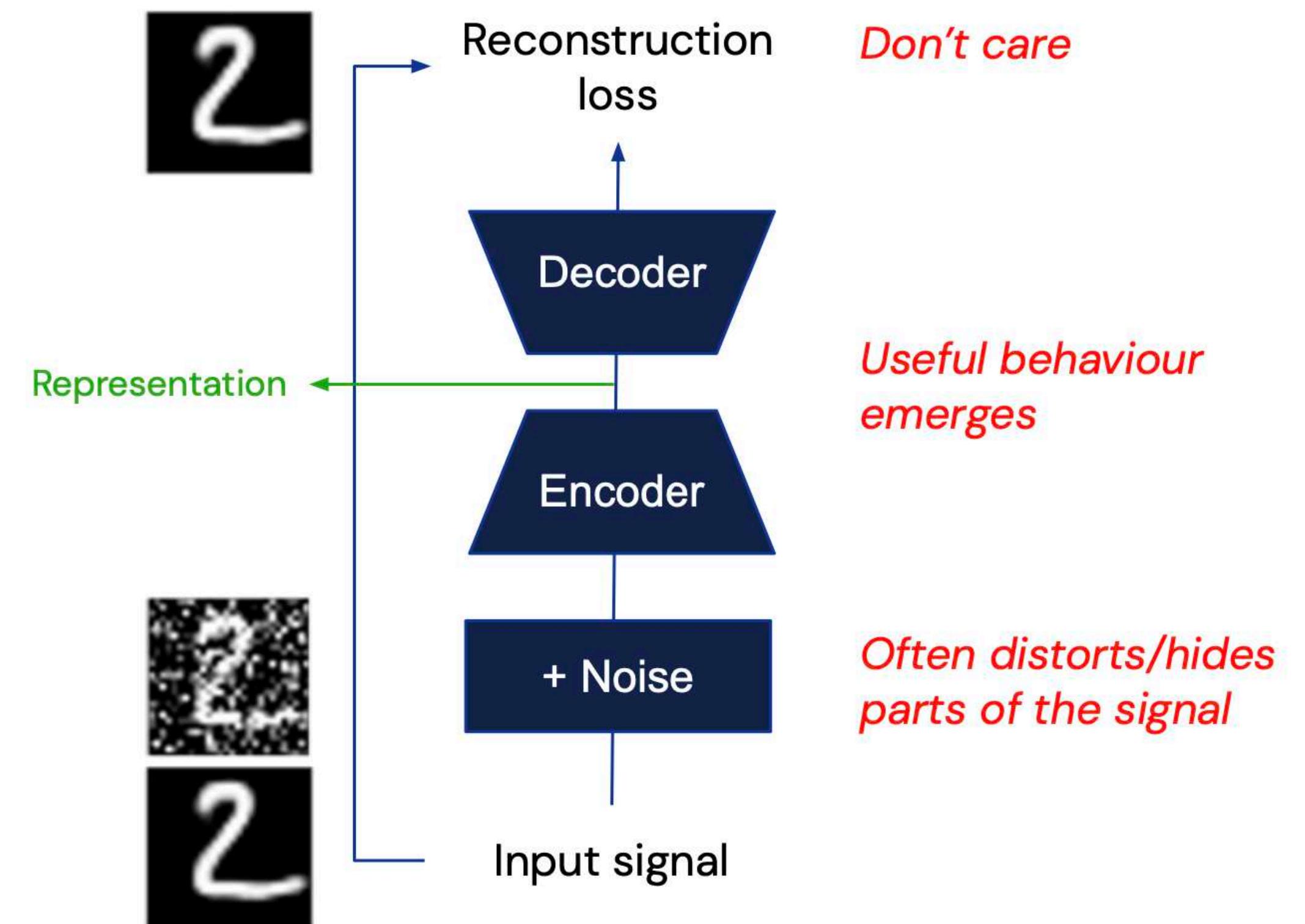
- A. Supervisado automático
- B. Tareas de pretexto
- C. Aprendizaje de representaciones
- D. Transferencia “downstream”



Aprendizaje Auto-supervisado

Self-Supervised Learning (SSL): Es un paradigma de entrenamiento:

- Objetivo de **aprender representaciones** de las imágenes
- Se usan datos no etiquetados
- A través de tareas auxiliares (pretexto) se crean pseudo-etiquetas usando la estructura de los datos

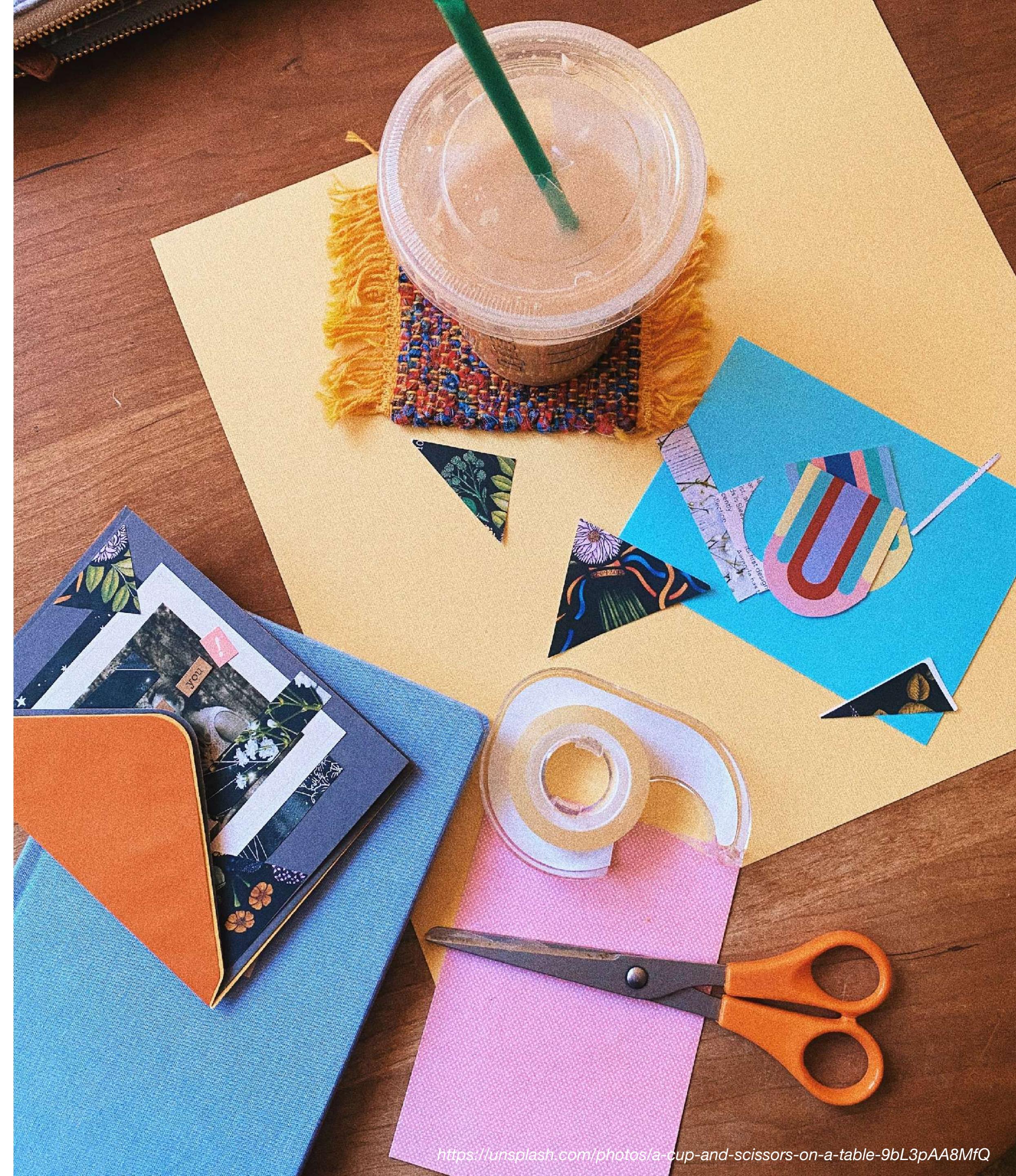


Tareas de pretexto

Principio básicos

Tareas artificiales diseñadas para alentar al modelo a aprender representaciones útiles de los datos.

Por ejemplo: Predecir las partes que faltan en una imagen



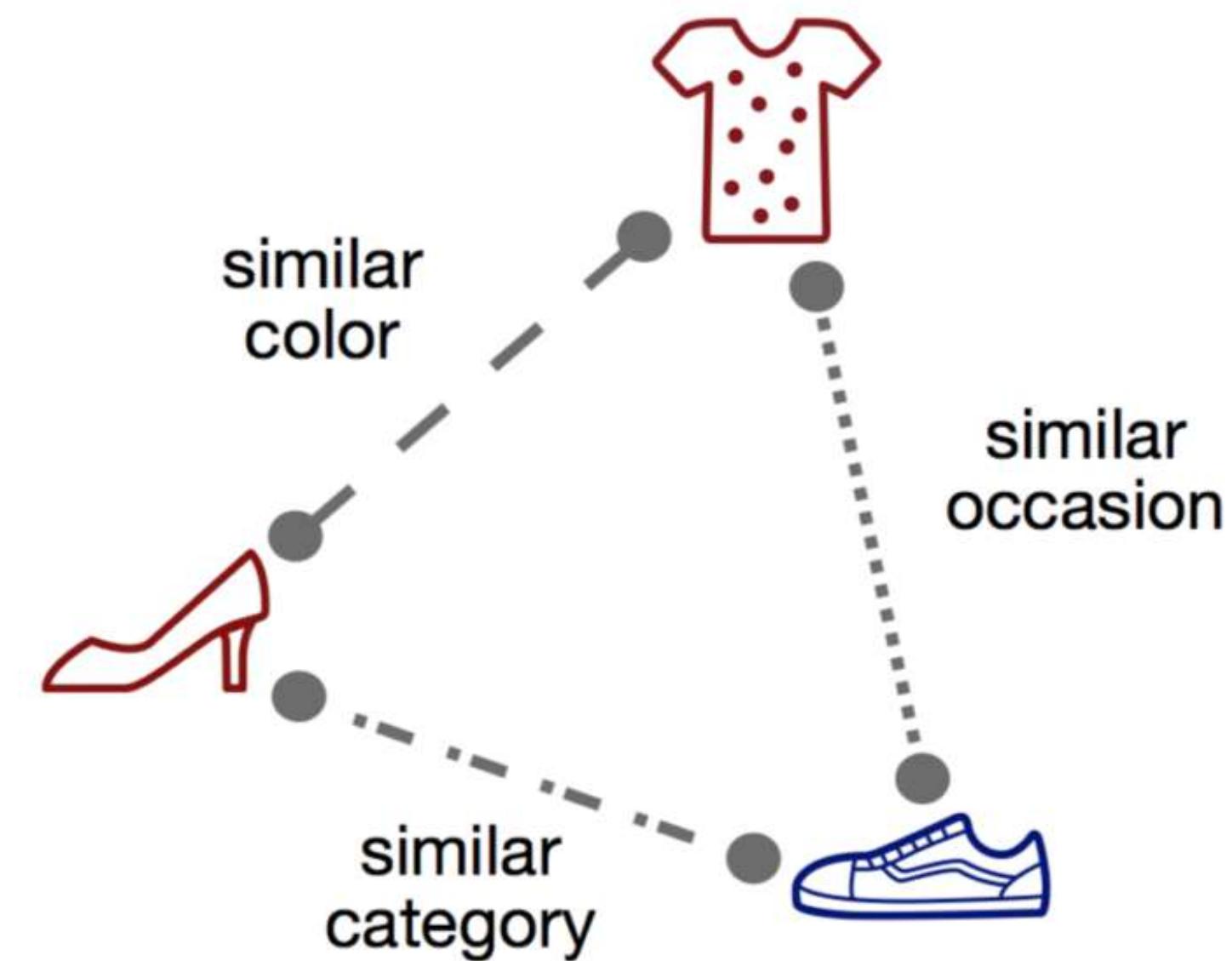
Representation Learning

Principio básicos

El objetivo es producir “**embeddings**”
(representaciones vectoriales).

Capturan información semántica. Ejemplo:
Formas, relaciones, sentido y significado
entre objetos.

Énfasis en el espacio de embebido como
representación del conocimiento



Downstream Transfer

Principio básicos

Después del preentrenamiento:

Las representaciones aprendidas son usadas para tareas específicas.

Ó el modelo se utiliza directamente para el aprendizaje tipo “zero-shot”.



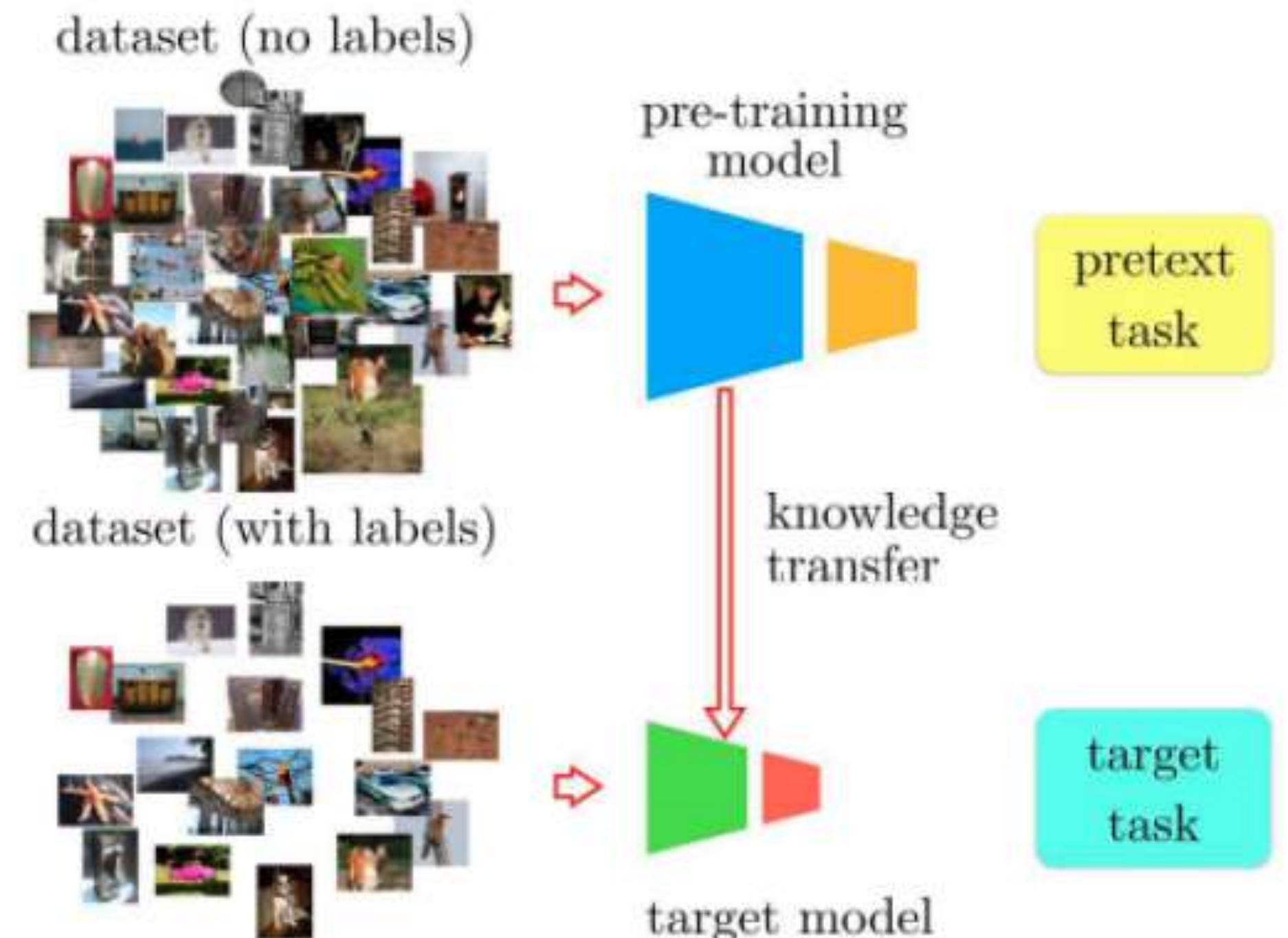
¿Por qué?

Aprendizaje Auto-supervisado

Abundancia de datos: Los datos no etiquetados son abundantes y más baratos de recopilar en comparación con los conjuntos de datos etiquetados

Generalización: SSL enseña a los modelos a extraer representaciones aplicables a diversas tareas “downstream”

Eficiencia: Reduce la necesidad de datos anotados específicos de la tarea durante el entrenamiento



SSL y ViT

- La SSL es usado especialmente en ViT pero no son exclusivos uno del otro
- Los ViT necesitan grandes cantidades de datos para aprender representaciones eficaces
- SSL proporciona una vía para el pre-entrenamiento en conjuntos de datos a gran escala

Aproximaciones

Aprendizaje Auto-supervisado

Las aproximaciones a SSL generalmente se clasifican en:

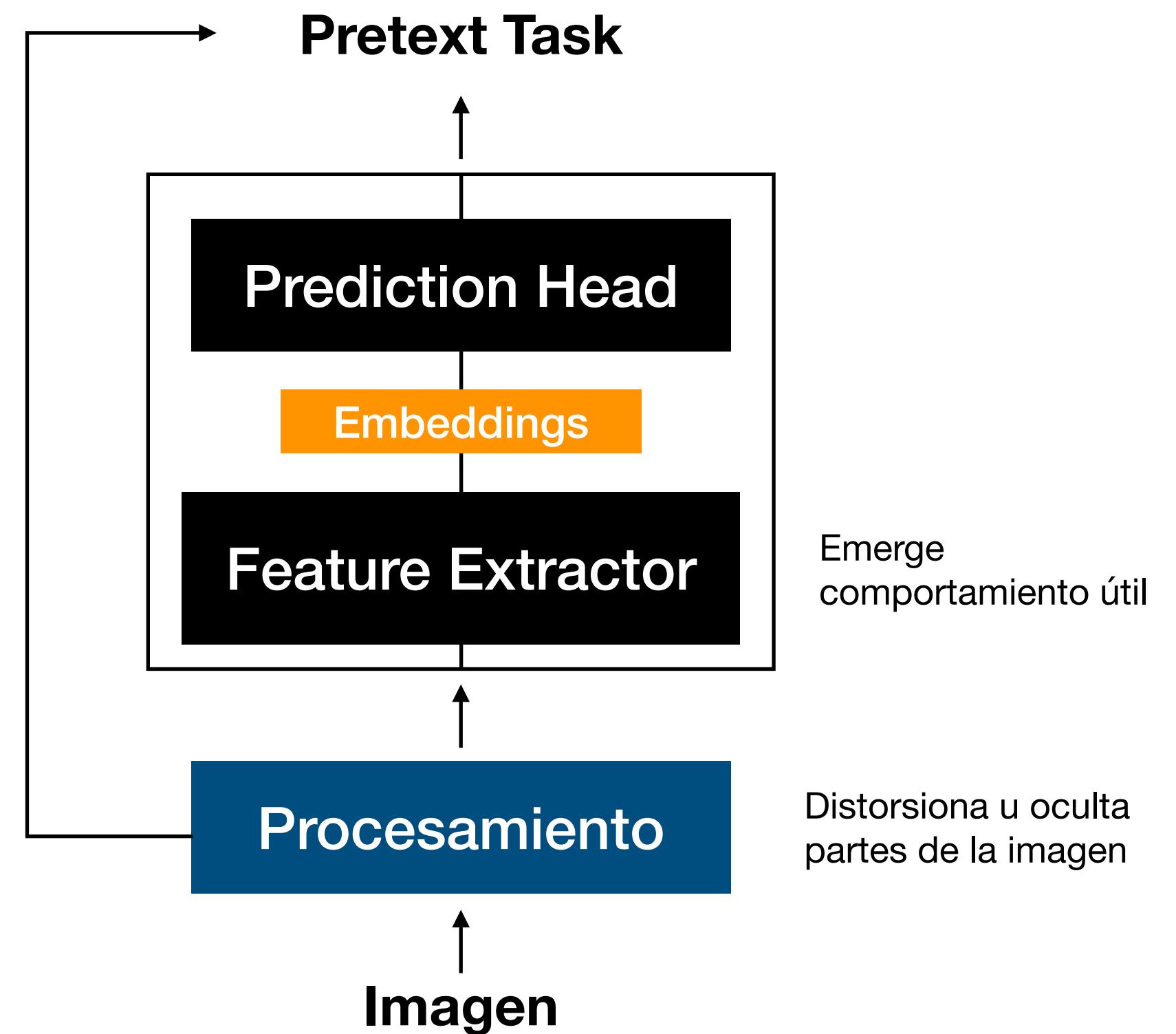
1. Predictivos
2. Contrastivos
3. Generativos



Agenda

- Modelos fundacionales
- Self-Supervised Learning
 - **Predictivos - Pretext tasks**
 - Contrastivos (SimSLR)
 - Generativos (MAE)
 - Discusión
 - Taller





Métodos predictivos

Tarea de Pretexto

- Predicir propiedades intrínsecas o transformaciones de los datos
- No importa la tarea como tal
- Solo importa que facilita el aprendizaje

Tareas Pretexto

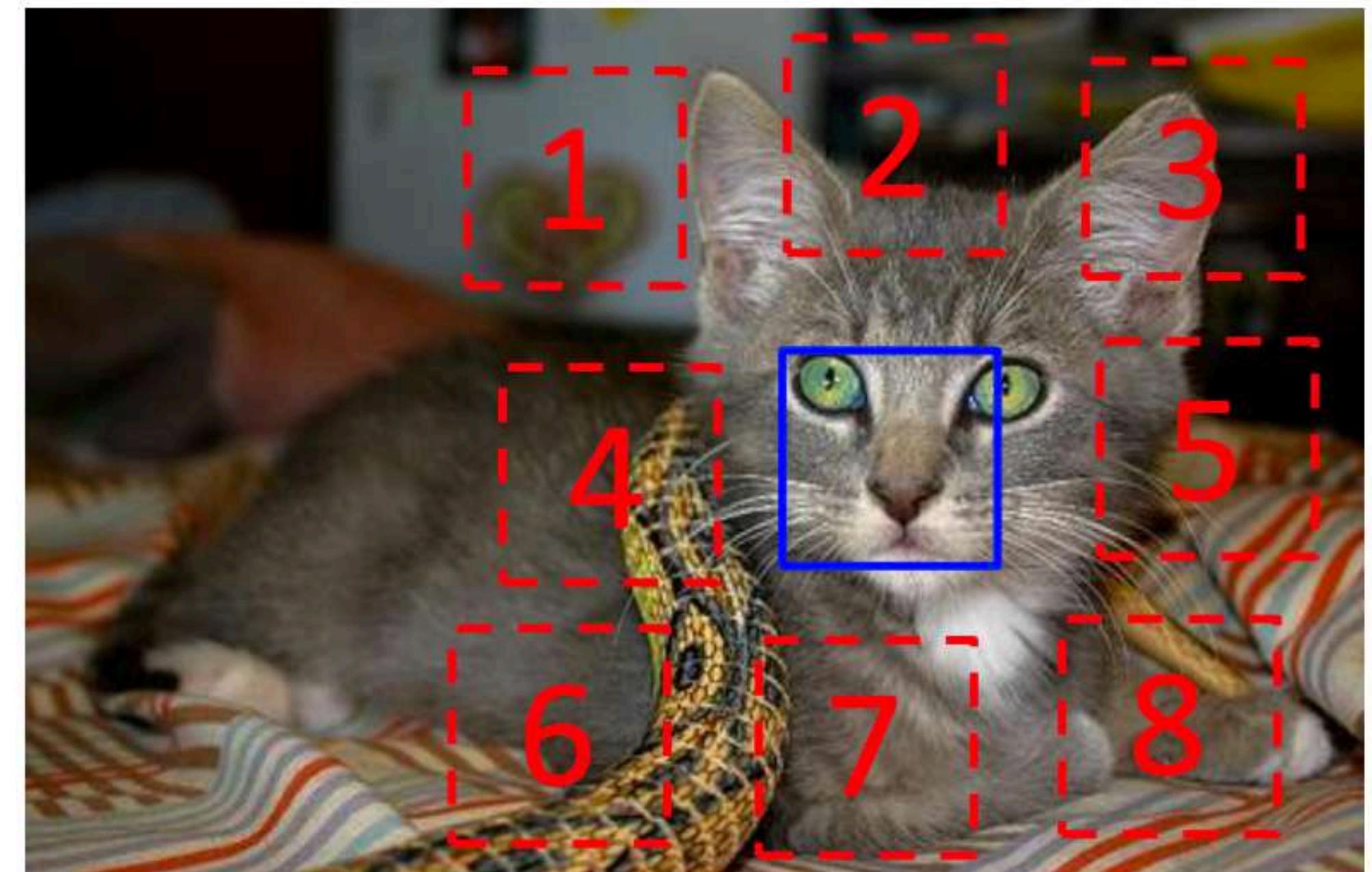
Algunos tipos de tareas pretexto:

- Inferencia de estructura
- Predicción de transformación
- Reconstrucción
- Temporales
- Multimodales
- Clasificación de instancias

Tareas Pretexto

Predicción de contexto

- Uno de los primeros métodos de SSL: Se extrae un parche aleatorio de una imagen y se utiliza como posición central
- A partir del parche central, podemos extraer 1 de otros 8 parches (en forma de cuadrícula con cierta fluctuación)
- Dado un par de parches como entrada: la red se entrena para aprender la posición relativa del parche central con respecto a su vecino.

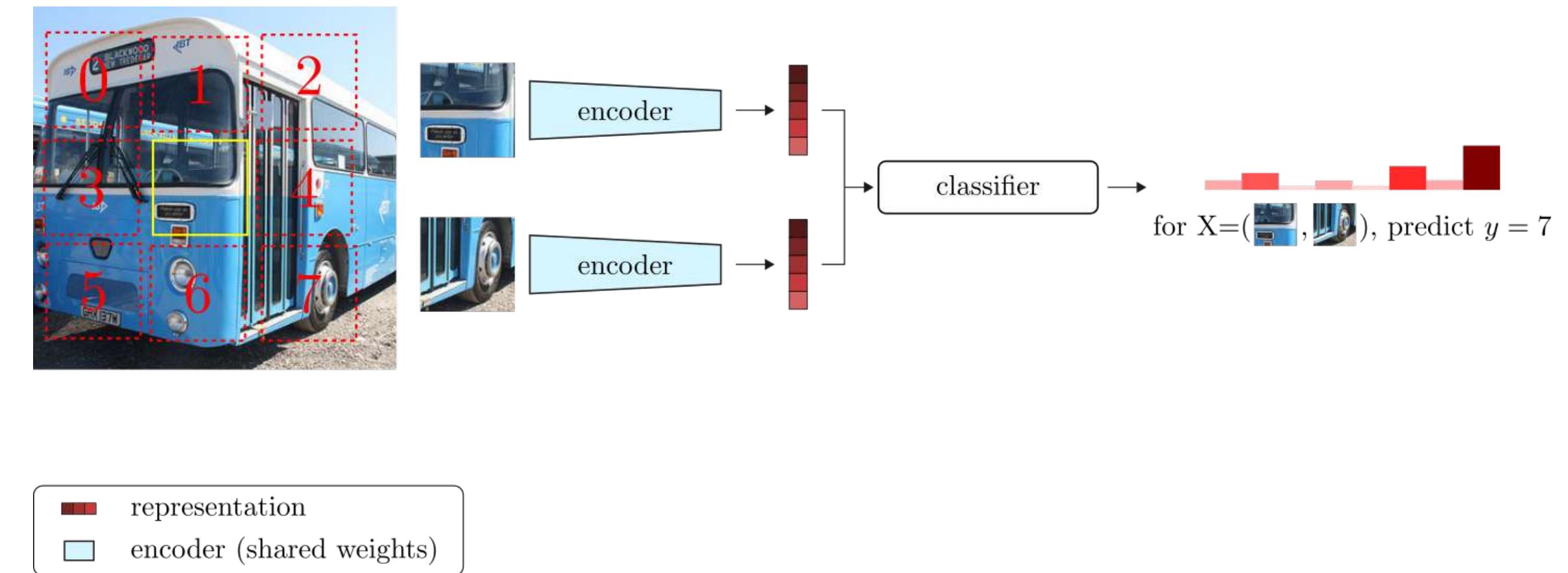


$$X = (\text{central patch}, \text{patch 3}); Y = 3$$

Tareas Pretexto

Predicción del contexto

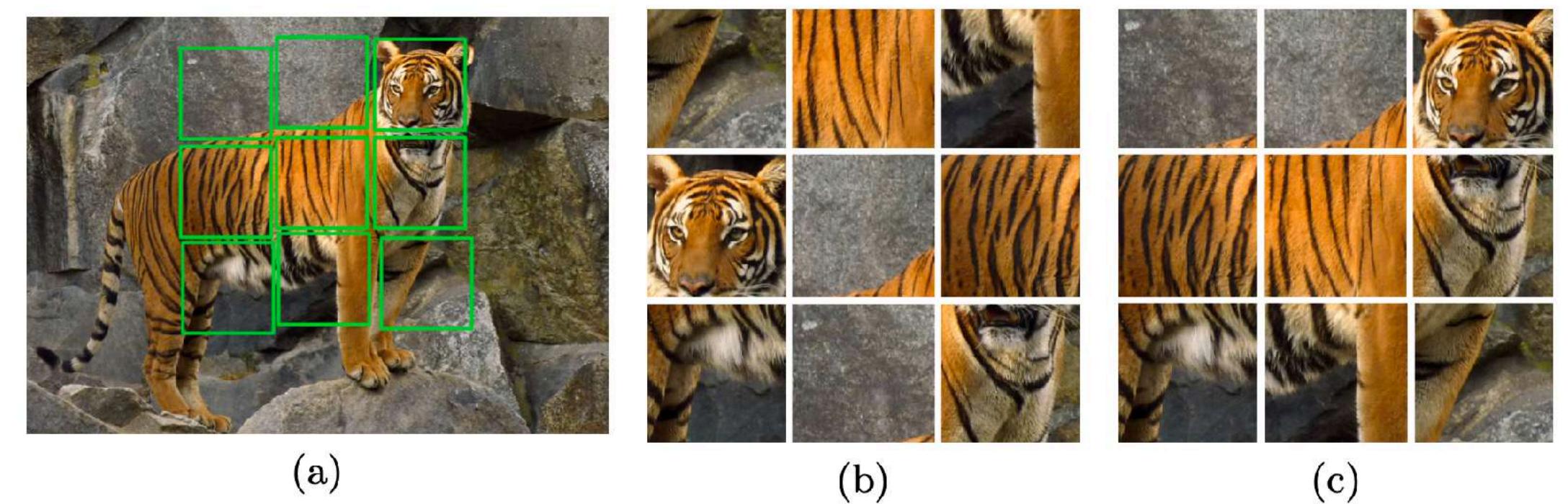
- Problema de clasificación de 8 categorías utilizando (log-loss)
- Motiva al aprendizaje de la estructura
- Tiene muchos problemas: No tiene suficiente variabilidad debido a que no hay negativos (parches de otras imágenes)
- Espacio de salida muy limitado: solo 8 posiciones para distinguir
- Diferencia muy grande con el uso en predicción



- Estudios en psicología muestran que los rompecabezas pueden utilizarse para evaluar el procesamiento visoespacial en humanos
- Propone utilizar rompecabezas para desarrollar una representación visoespacial de objetos en el contexto de las CNN

Tareas Pretexto

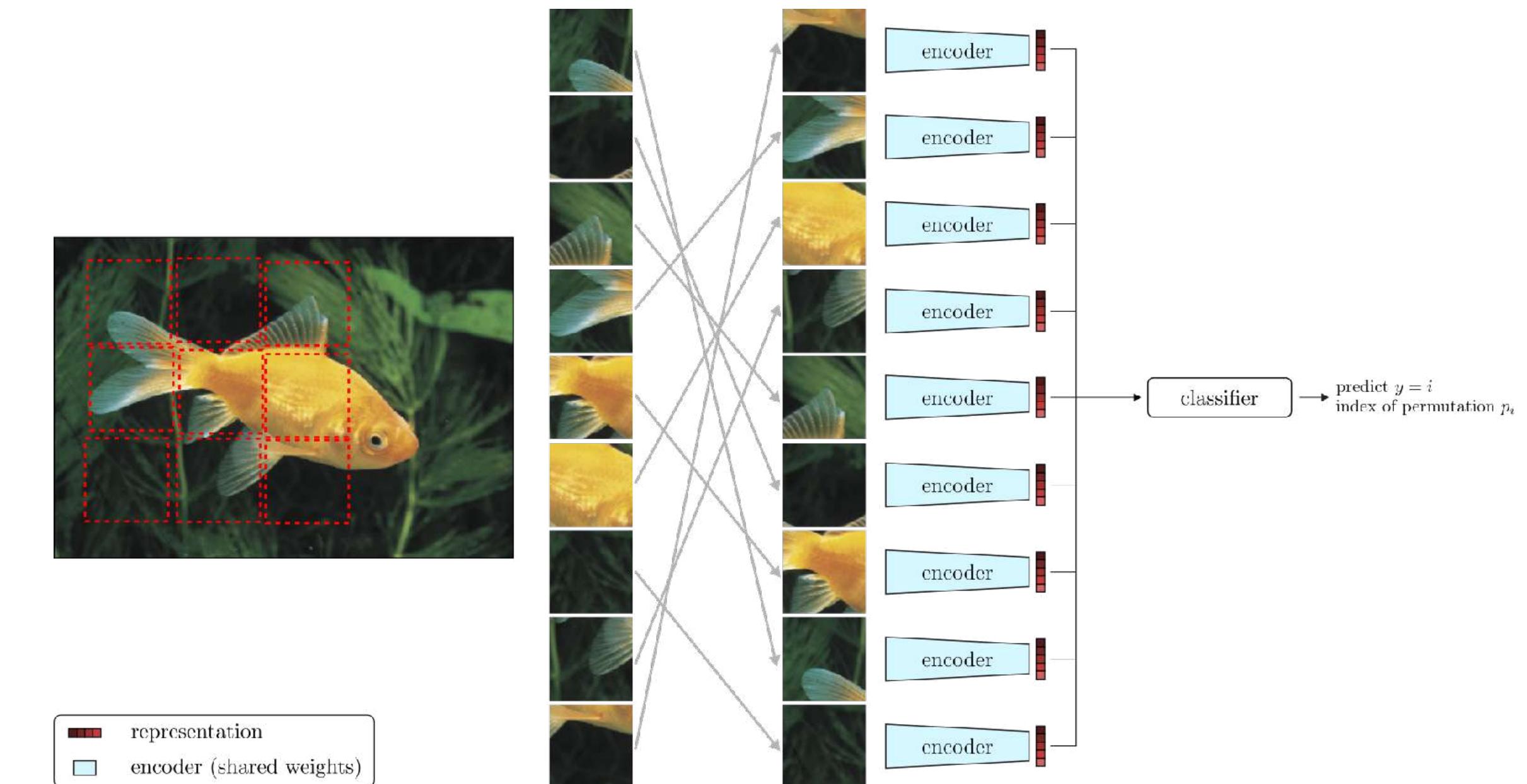
Jigsaw Puzzle



Tareas Pretexto

Jigsaw Puzzle

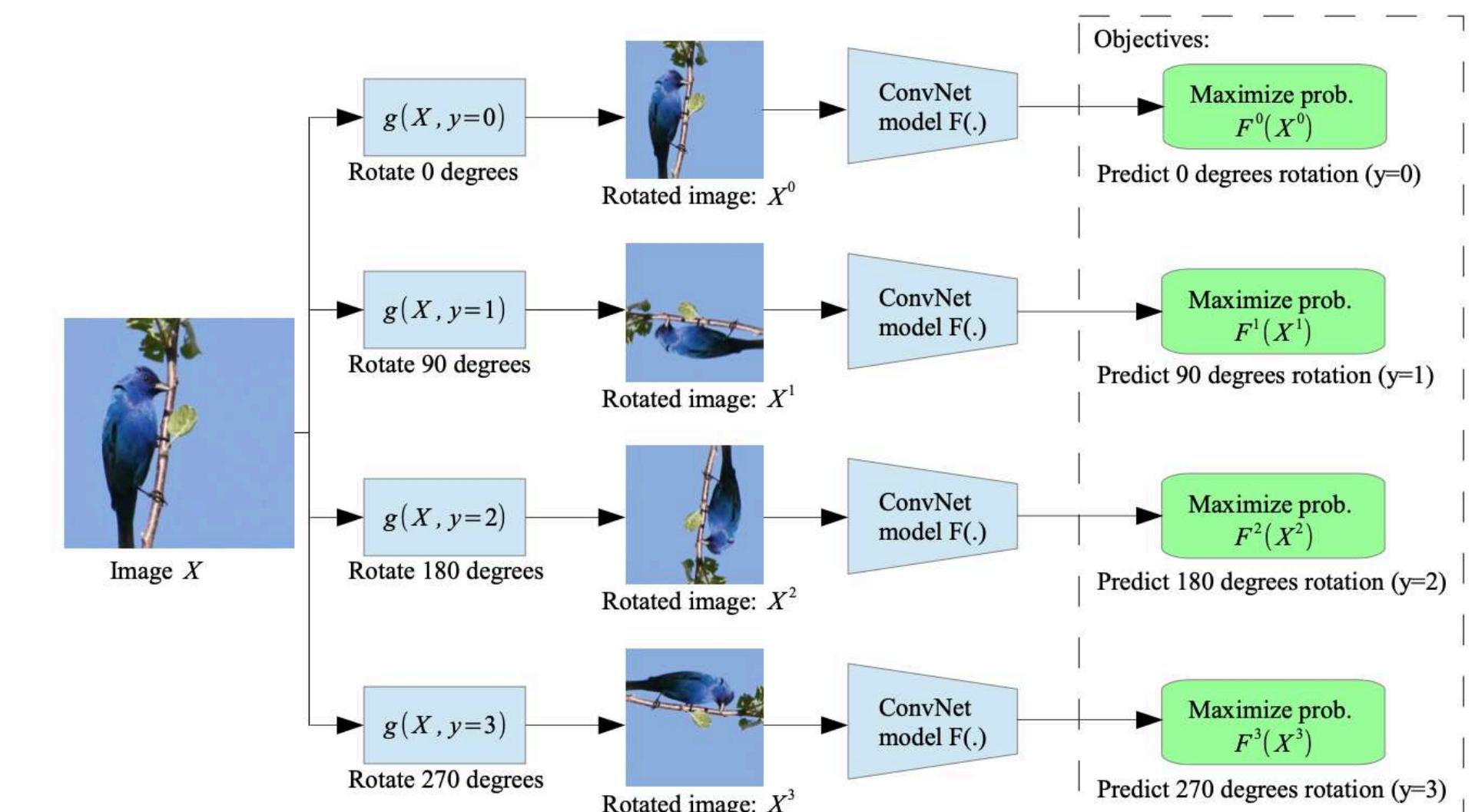
- Definir un conjunto de permutaciones del rompecabezas (Ej: $S = \{3,1,2,9,5,4,8,7,6\}$) y asignar un índice a cada permutación
- Recortar una área de la imagen de forma aleatoria y dividir en 3×3 con fluctuaciones
- Elige aleatoriamente una de esas permutaciones, reordena los 9 parches de entrada según esa permutación
- Pedir al modelo que devuelva un vector con el valor de probabilidad del índice



Tareas Pretexto

Rotación

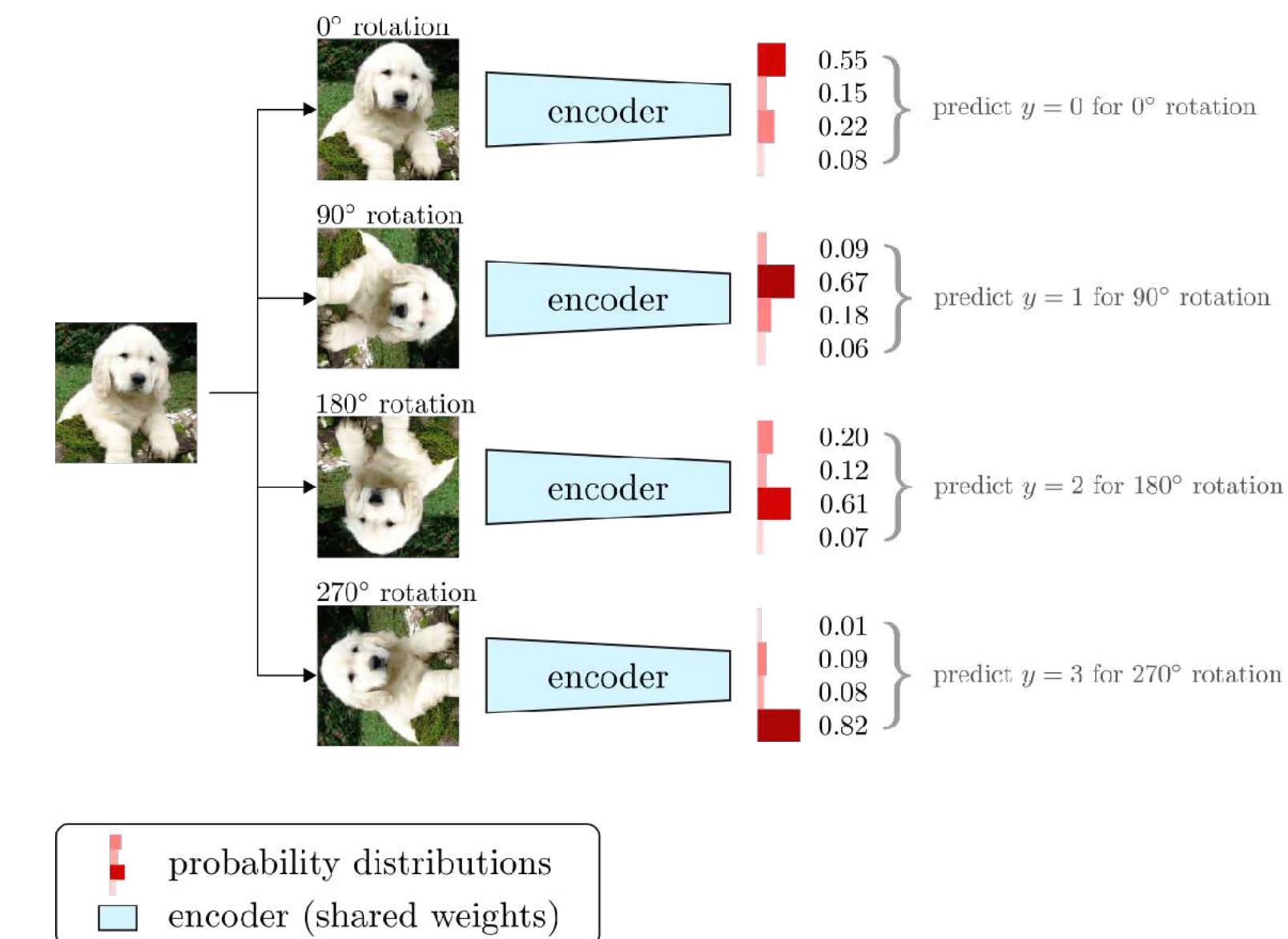
- Predicción del ángulo de rotación aplicado a una imagen de entrada
- Para cada imagen de entrada se define un ángulo de rotación β (elegido aleatoriamente entre un conjunto de valores predefinidos)
- La imagen se rota un ángulo y se introduce como entrada en el modelo



Tareas Pretexto

Rotación

- Implementación muy simple y efectiva
- No es lo suficientemente genérico debido a la ausencia de negativos de otras imágenes (no hay razón para distinguir el gato del perro)
- Pequeño espacio de salida - 4 casos (rotaciones)
- Hay dominios más fáciles que otros: (calles o playas)



Tareas Pretexto

Bag-of-Words

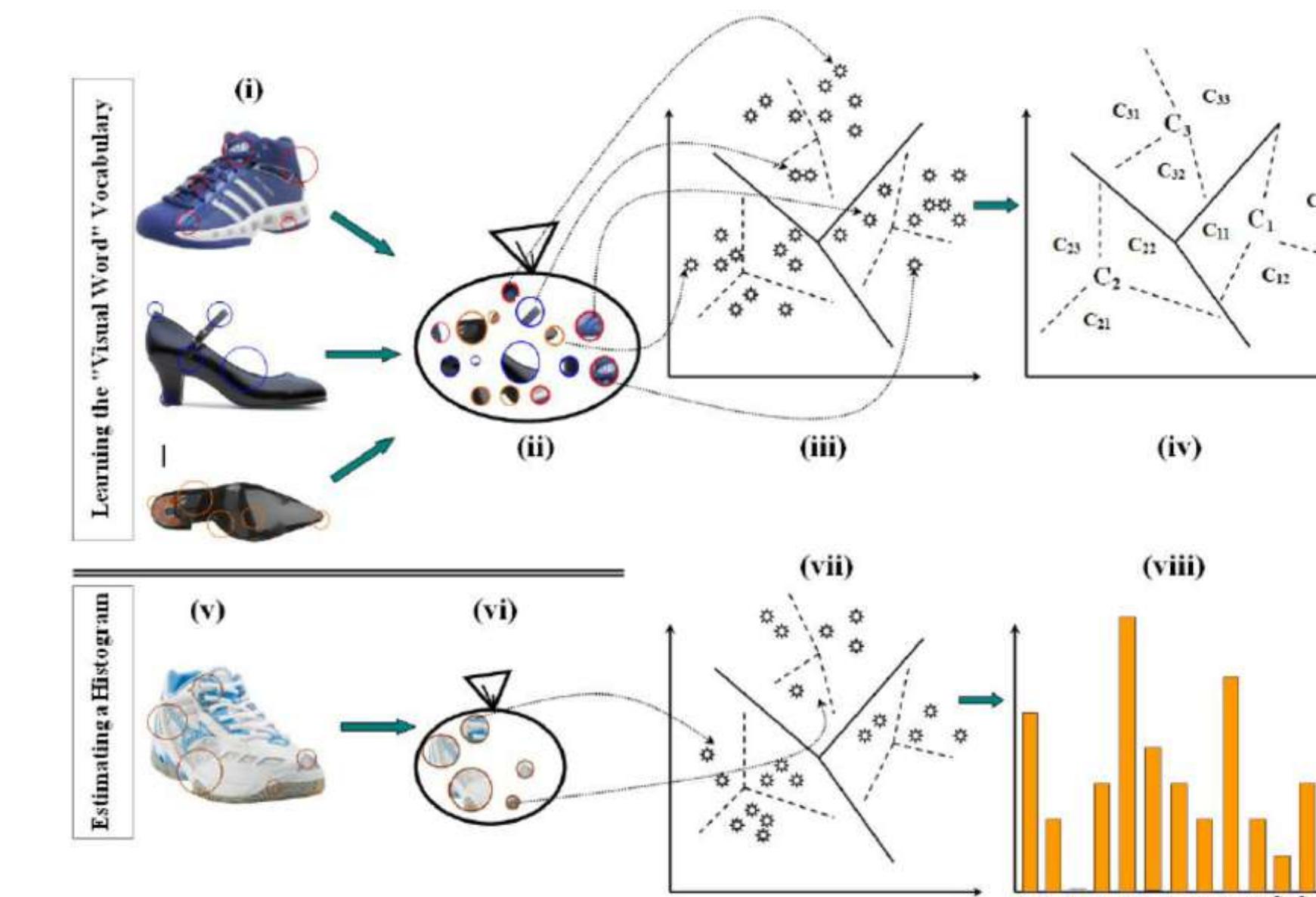
- Es una técnica de visión por computador tradicional inspirada en NLP popularmente usado para indexar y recuperar documentos
- Se presenta un documento como un histograma de palabras. Se ignora la gramática y el orden de las palabras
- Por el número de tipos de palabras que hay en la bolsa puedo saber ese documento a que categoría pertenece



Tareas Pretexto

Visual-Bag-of-Words

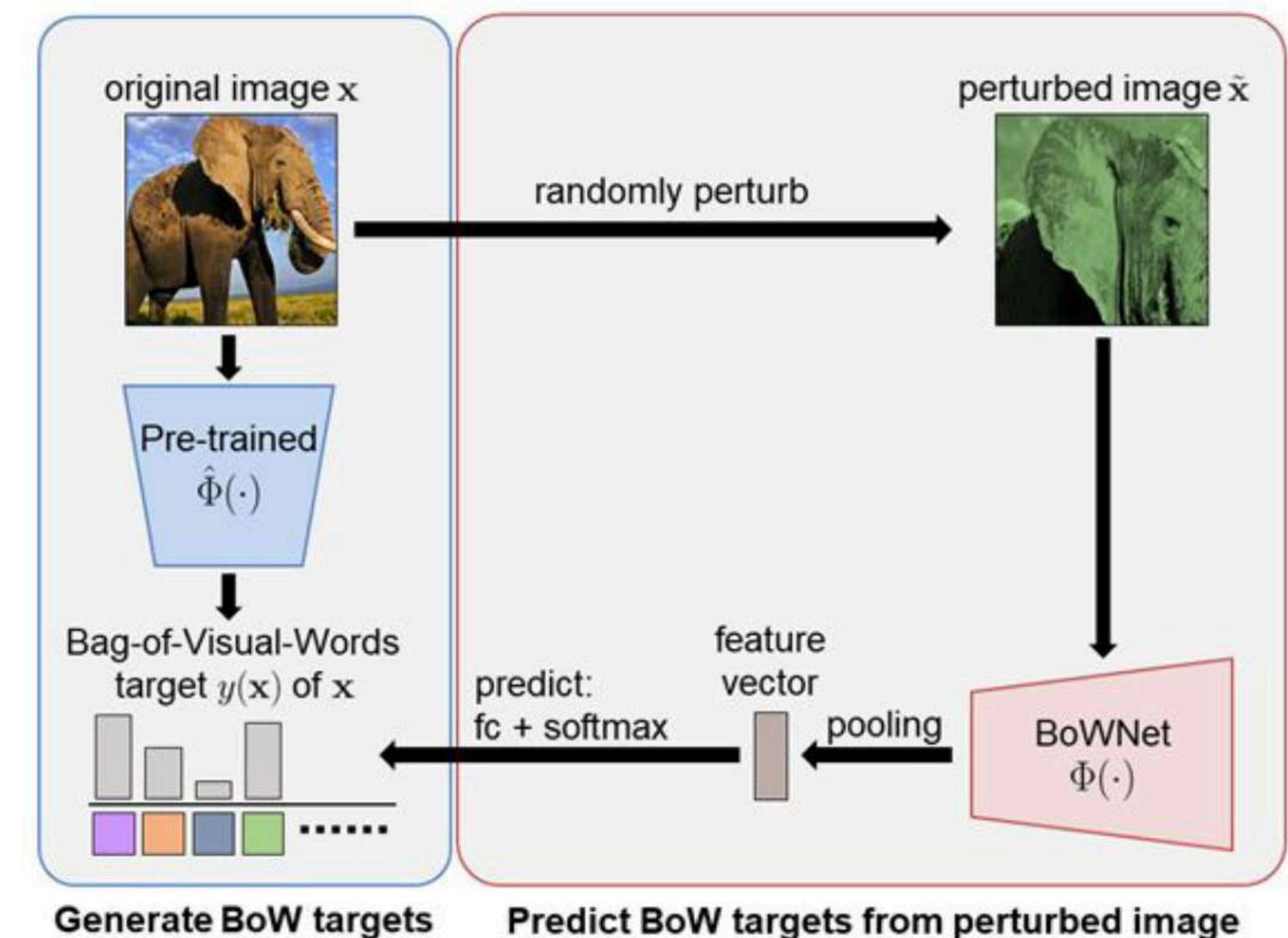
- Se extraen las descriptores de todo conjunto de datos
- Se crea un “codebook” (vocabulario) agrupando los descriptores con alguna técnica de clustering
- Cada imagen se representa por el conteo de sus descriptores a los clusters
- Este histograma es el vector de características que se usa con cualquier técnica de ML



Tareas Pretexto

Visual-Bag-of-Words

- Parte de una red pre-entrenada para extraer características y construir un “vocabulario”
- Se calcula el BoW de una imagen de entrenamiento con el modelo pre-entrenado
- La imagen original es distorsionada
- Se entrena otro modelo para predecir las BoW del modelo pre-entrenado con una función de costo “Cross Entropy Loss”



Agenda



- Modelos fundamentales
- Self-Supervised Learning
 - Predictivos - Pretext tasks
 - **Contrastivos (SimSLR)**
 - Generativos (MAE)
- Discusión
- Taller

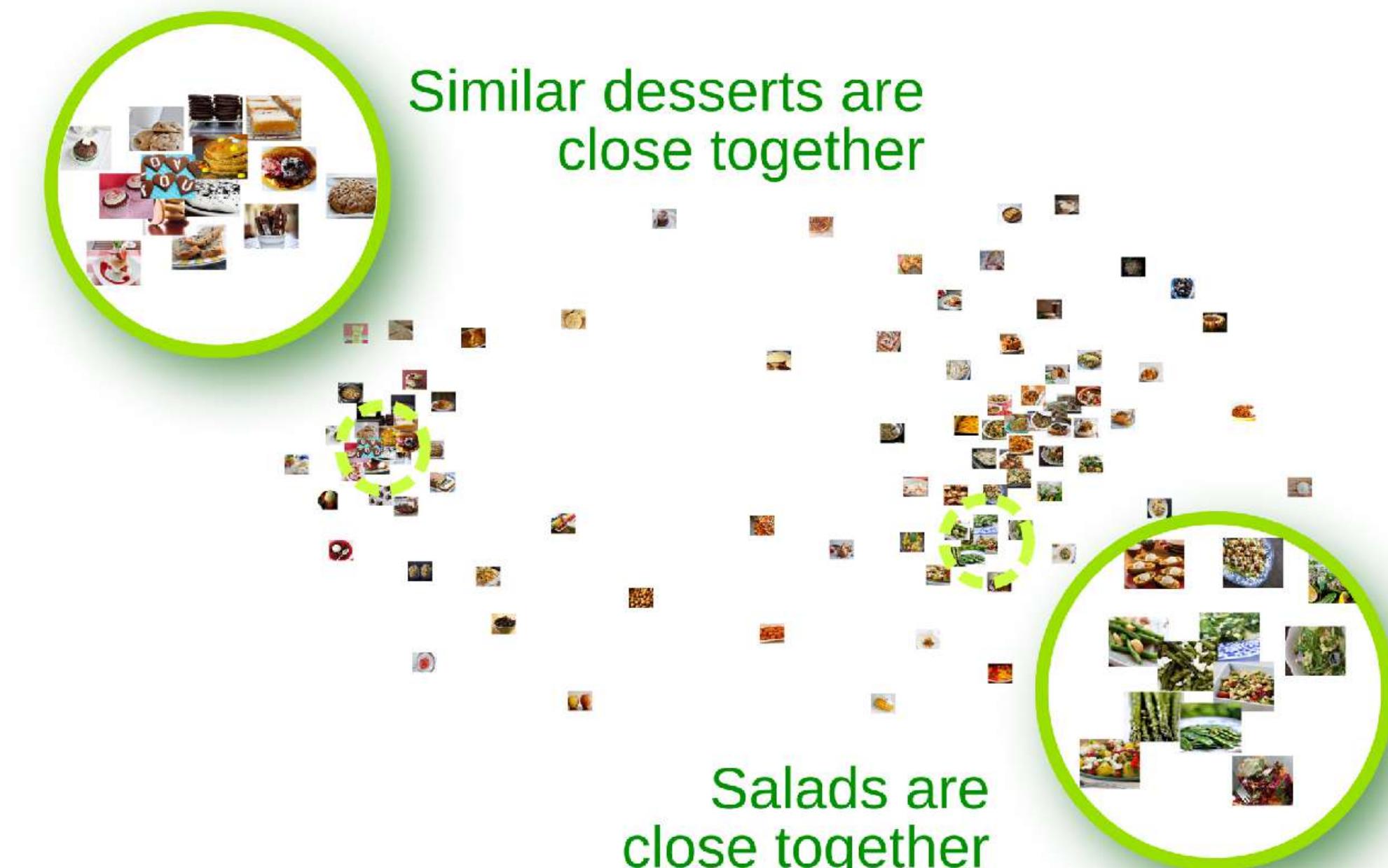
Contrastive Learning

Términos

Pares positivos: Muestras que son semántica o estructuralmente similares (por ejemplo, dos vistas aumentadas de la misma imagen)

Pares negativos: Muestras de datos que no son similares o no están relacionadas

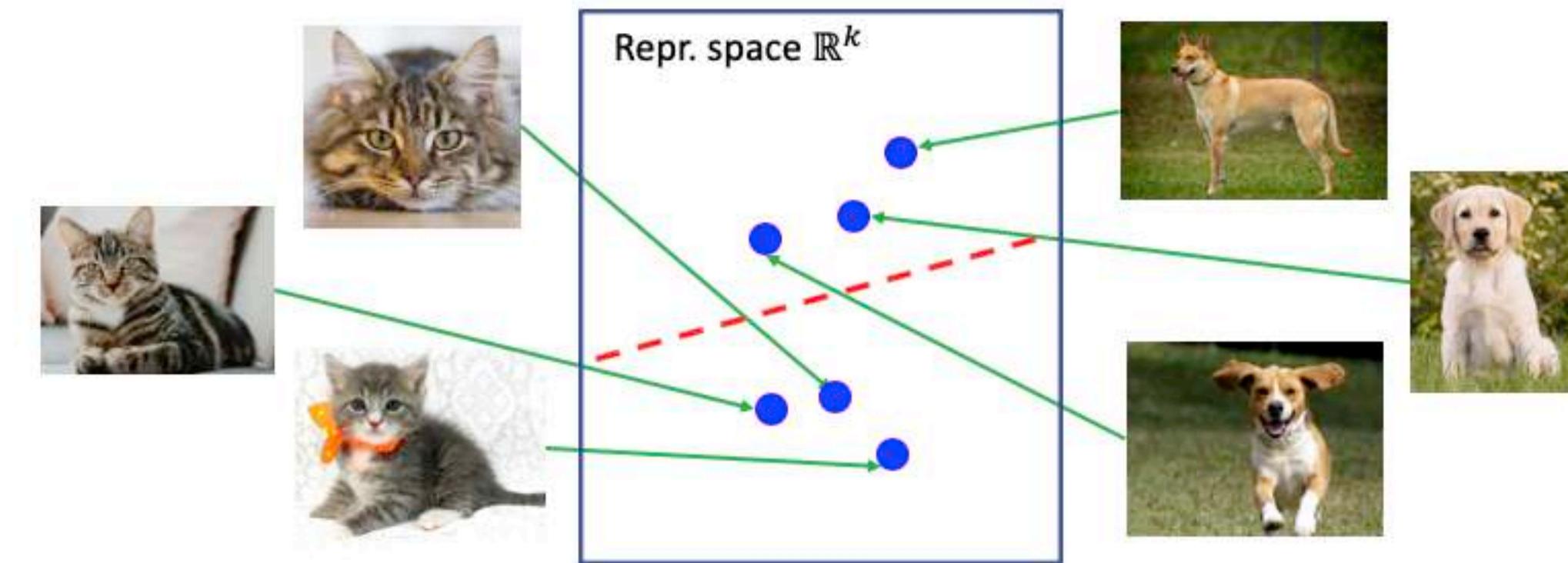
Embedding space: Espacio en el que se codifican las características extraídas por un modelo



Contrastivos

Aprendizaje Auto-supervisado

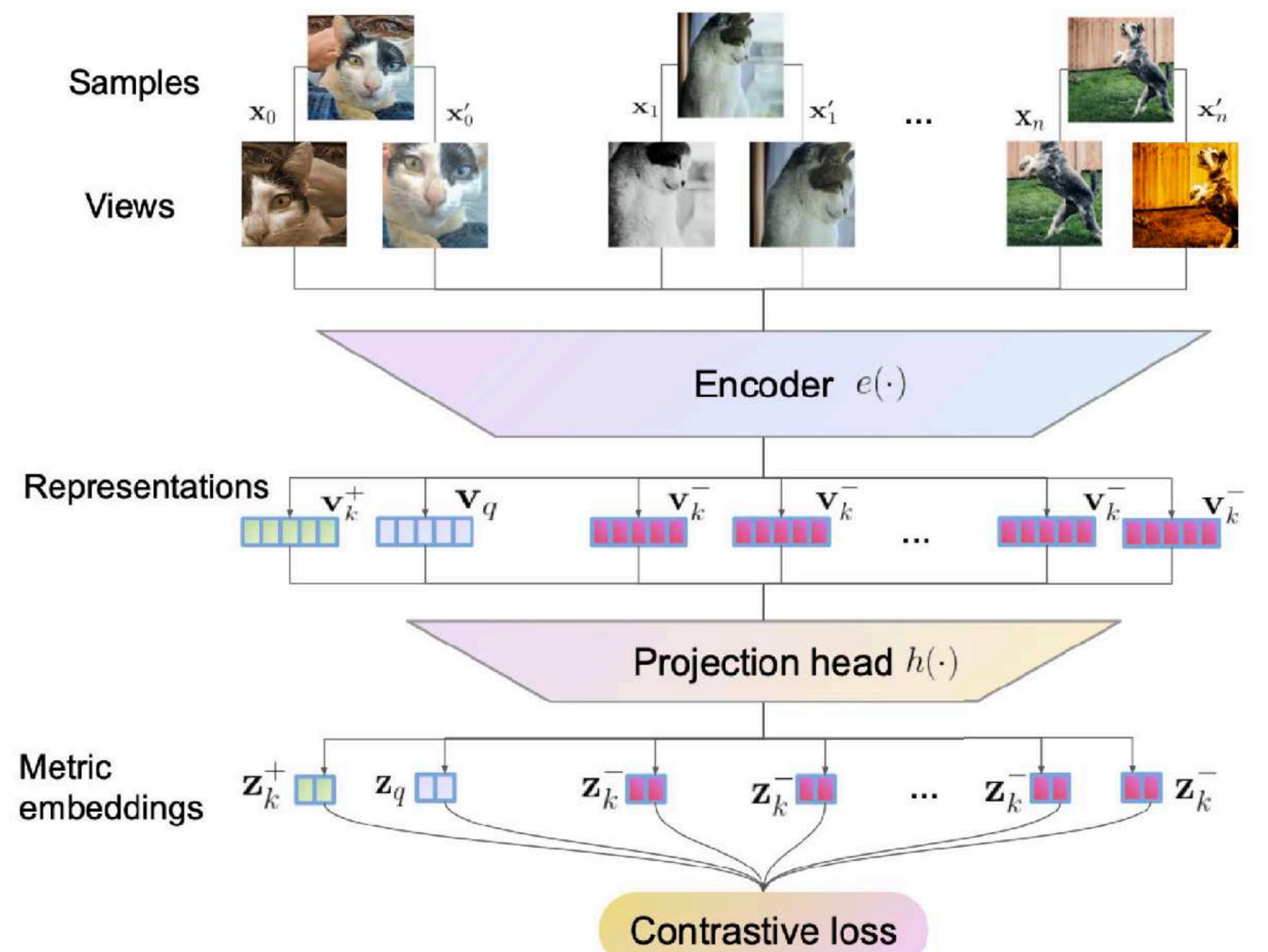
- El objetivo es aprender un espacio de embebidos en el que las muestras positivas permanezcan cerca y los negativos lejos
- En lugar de usar una pseudoetiqueta de la tarea de pretexto. Aprenden un modelo discriminativo sobre múltiples pares de entrada



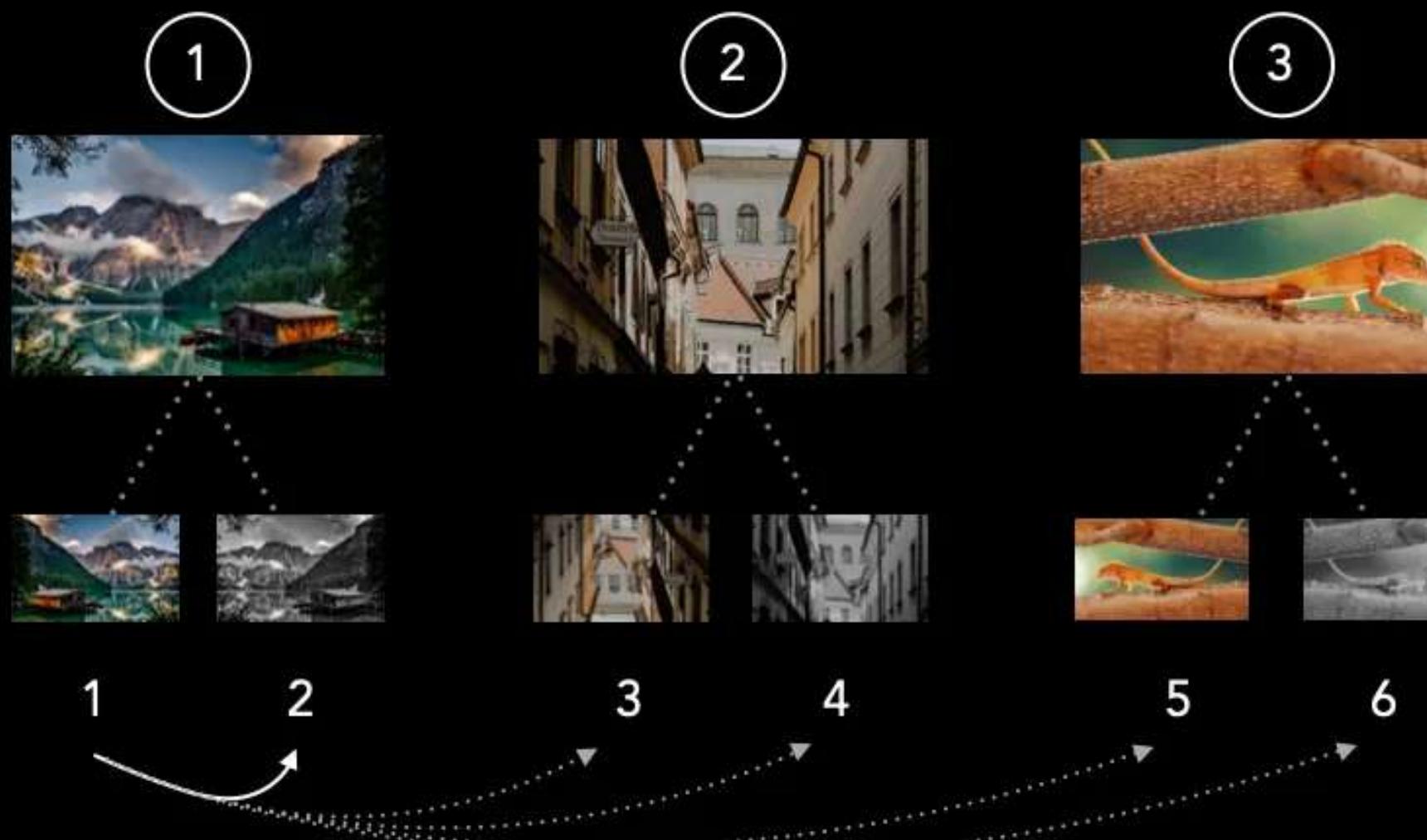
Contrastivos

SimCLR

1. Una imagen original se elige aleatoriamente
2. Se generan 2 vistas utilizando aumento de datos
3. Se calculan los embebidos con una CNN (basada en ResNet)
4. Realiza proyección con una NN
5. Se entrena para minimizar la función “contrastive loss”



Contrastive Loss



$$\mathcal{L} = -\frac{1}{N} \sum_i^N \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_j^K \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}$$

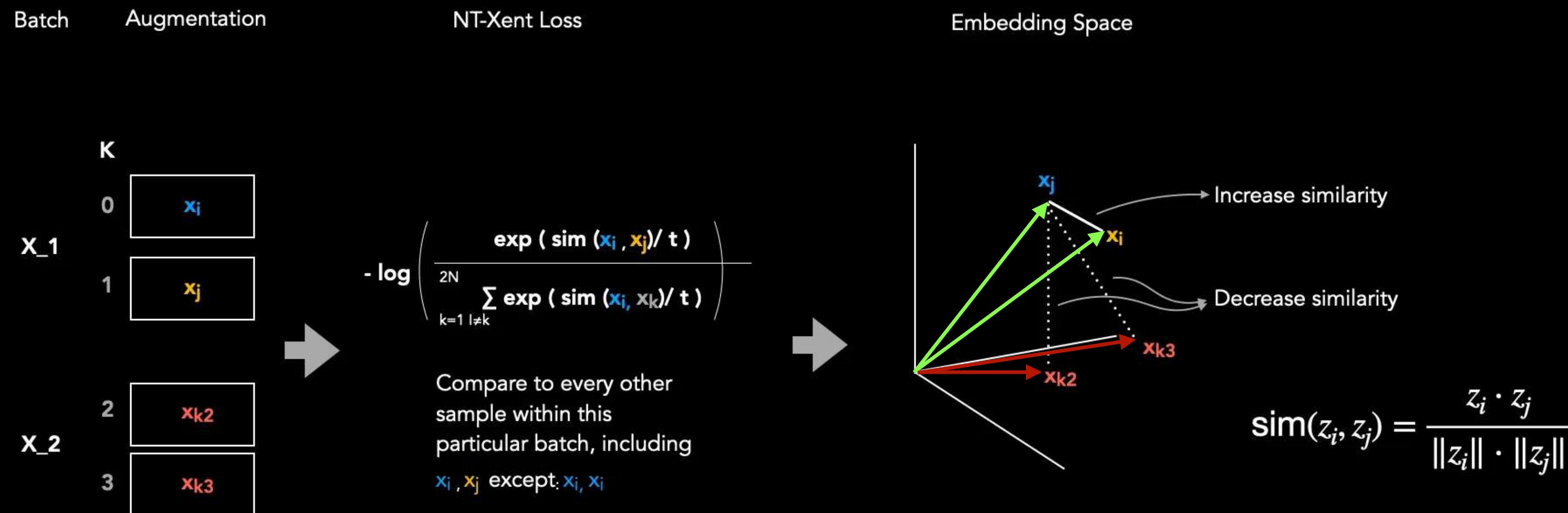
\downarrow

$$\frac{\text{sim}(1,2)}{\text{sim}(1,2)+\text{sim}(1,3) + \text{sim}(1,4) + \text{sim}(1,5) + \text{sim}(1,6)}$$

$$\frac{\mathbf{z}_i \cdot \mathbf{z}_i^+}{\|\mathbf{z}_i\| \|\mathbf{z}_i^+\|}$$

Cosine similarity

Contrastive Loss



Agenda



- Modelos fundamentales
- Self-Supervised Learning
 - Predictivos - Pretext tasks
 - Contrastivos (SimSLR)
 - **Generativos (MAE)**
- Discusión
- Taller

Generativos

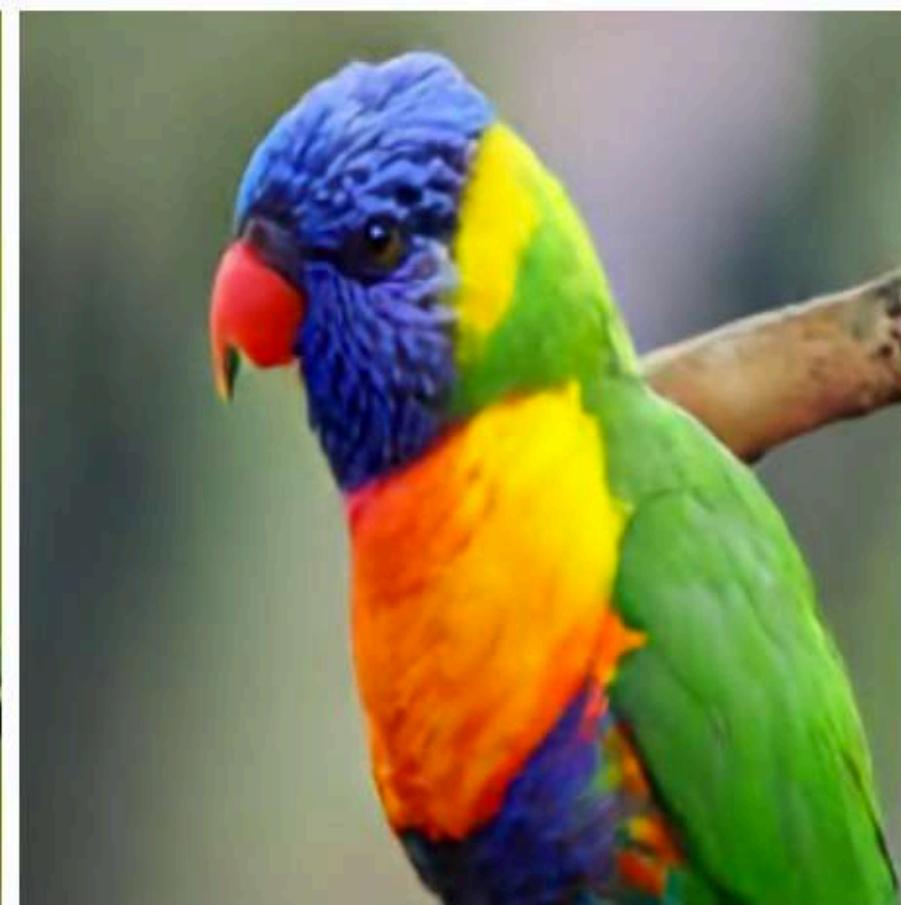
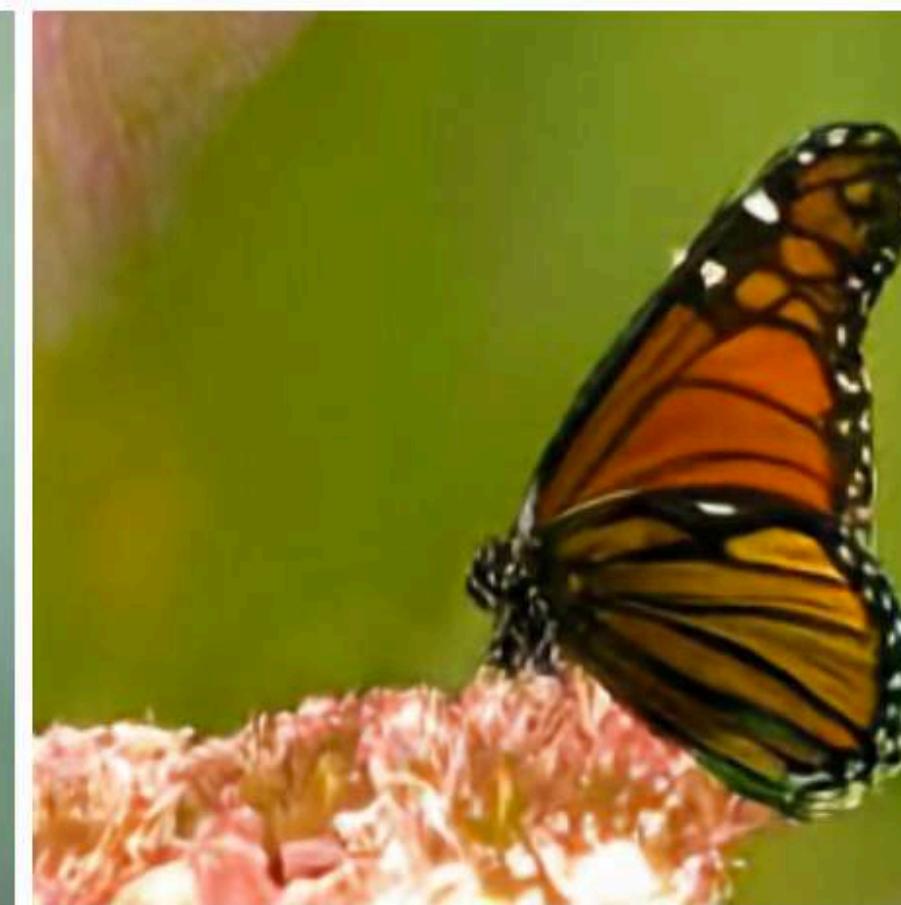
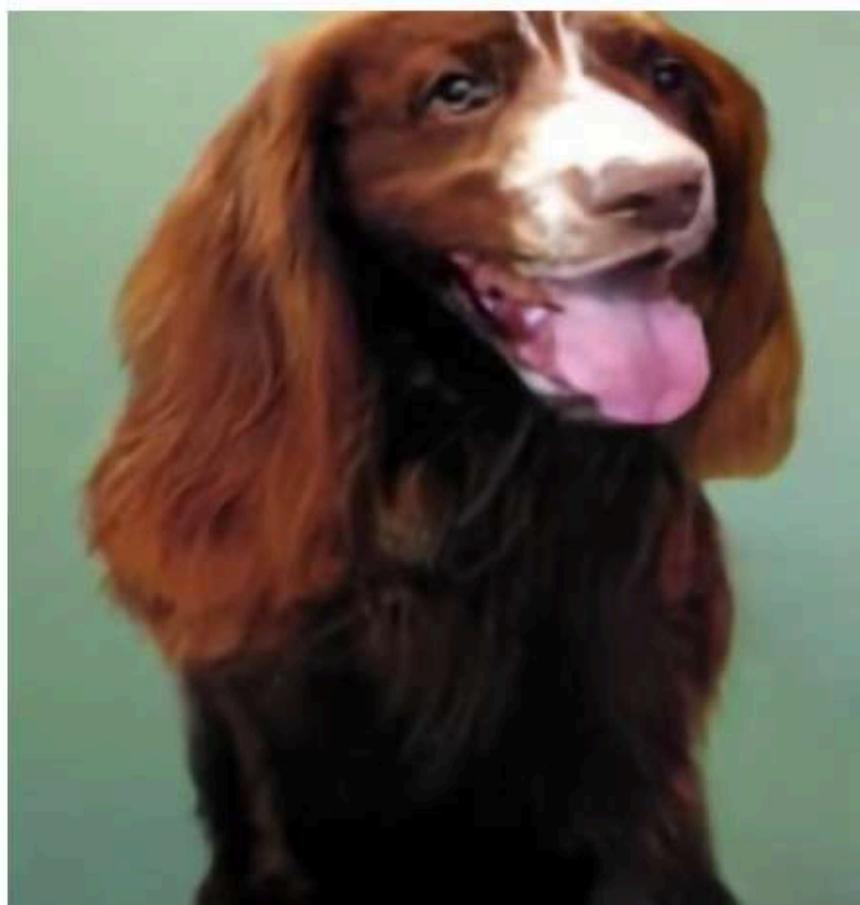
- Aprender una *representación latente* generando o reconstruyendo datos de entrada
- Dos aproximaciones:
 1. Modelar la distribución de los datos
 2. Reconstruir información faltante



Generativos

Modelar la distribución de los datos

Variational Autoencoders



VQ-VAE-2

Generative Adversarial Networks

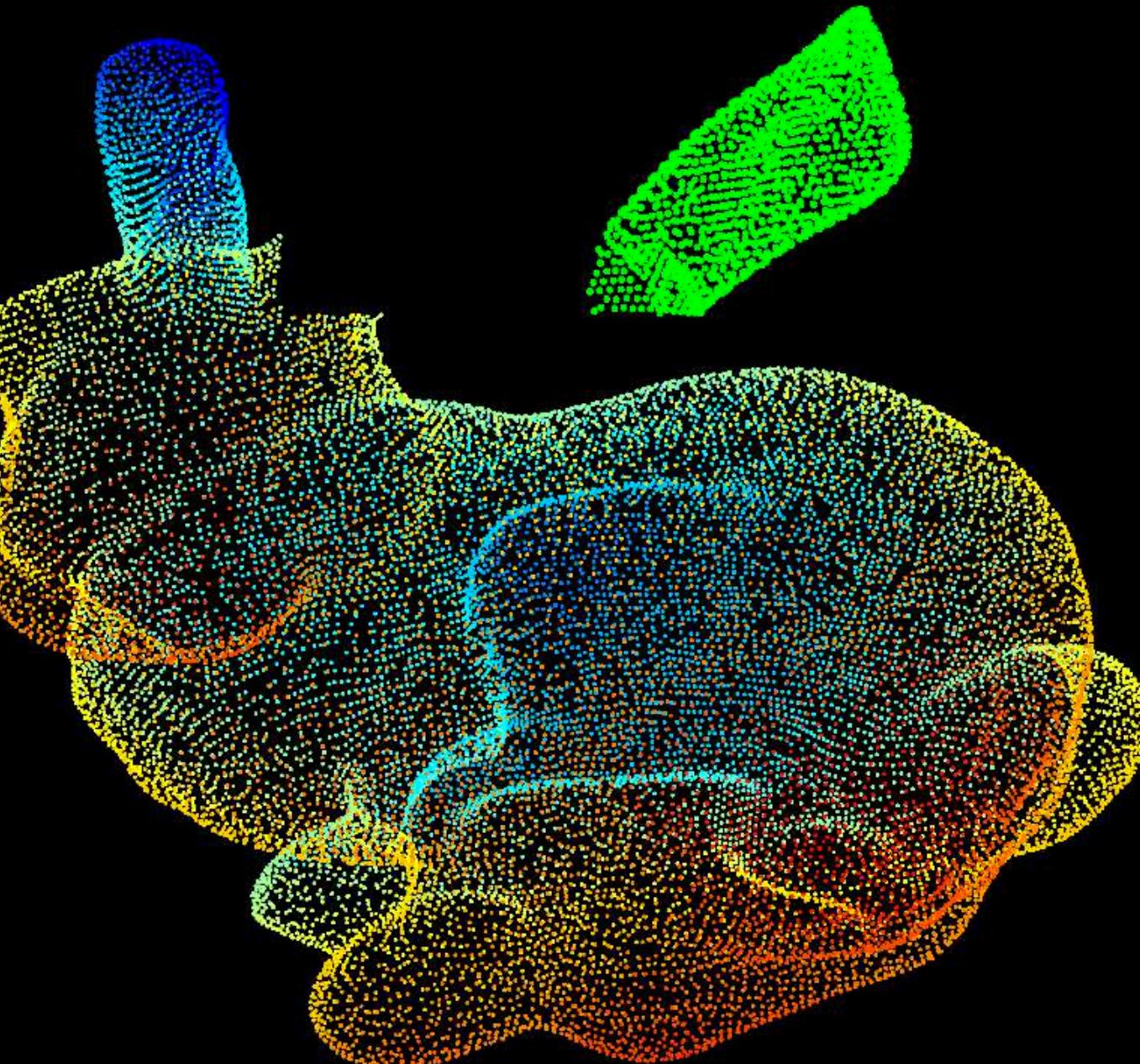


StyleGan-2

Generativos

2. Reconstruir información

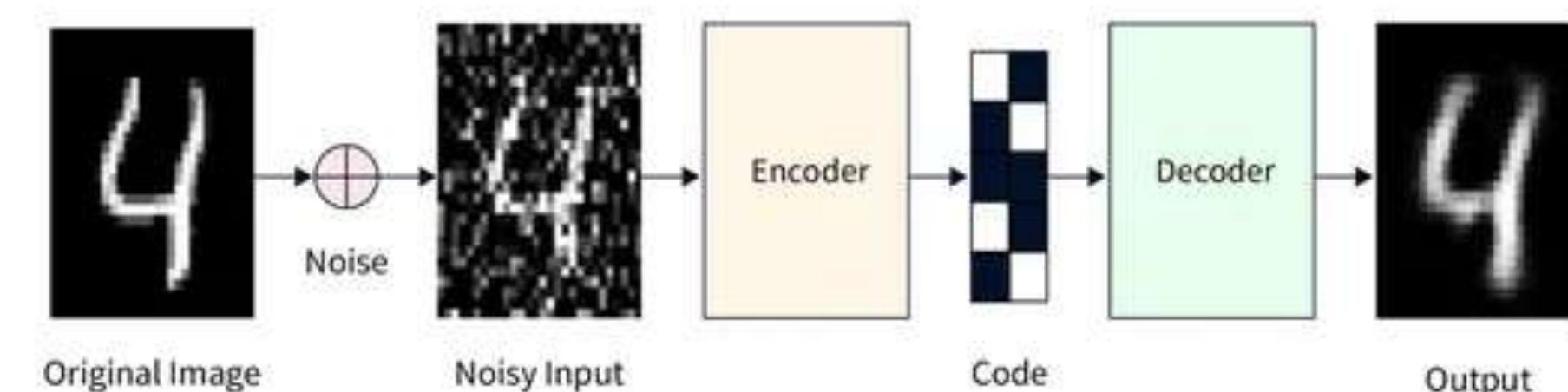
- Aprender representaciones prediciendo las partes que faltan de los datos de entrada
- Estos métodos aprovechan el contexto de las regiones visibles para inferir las partes que faltan
- Permite al modelo comprender la estructura y las relaciones de los datos



Generativos

Denoising

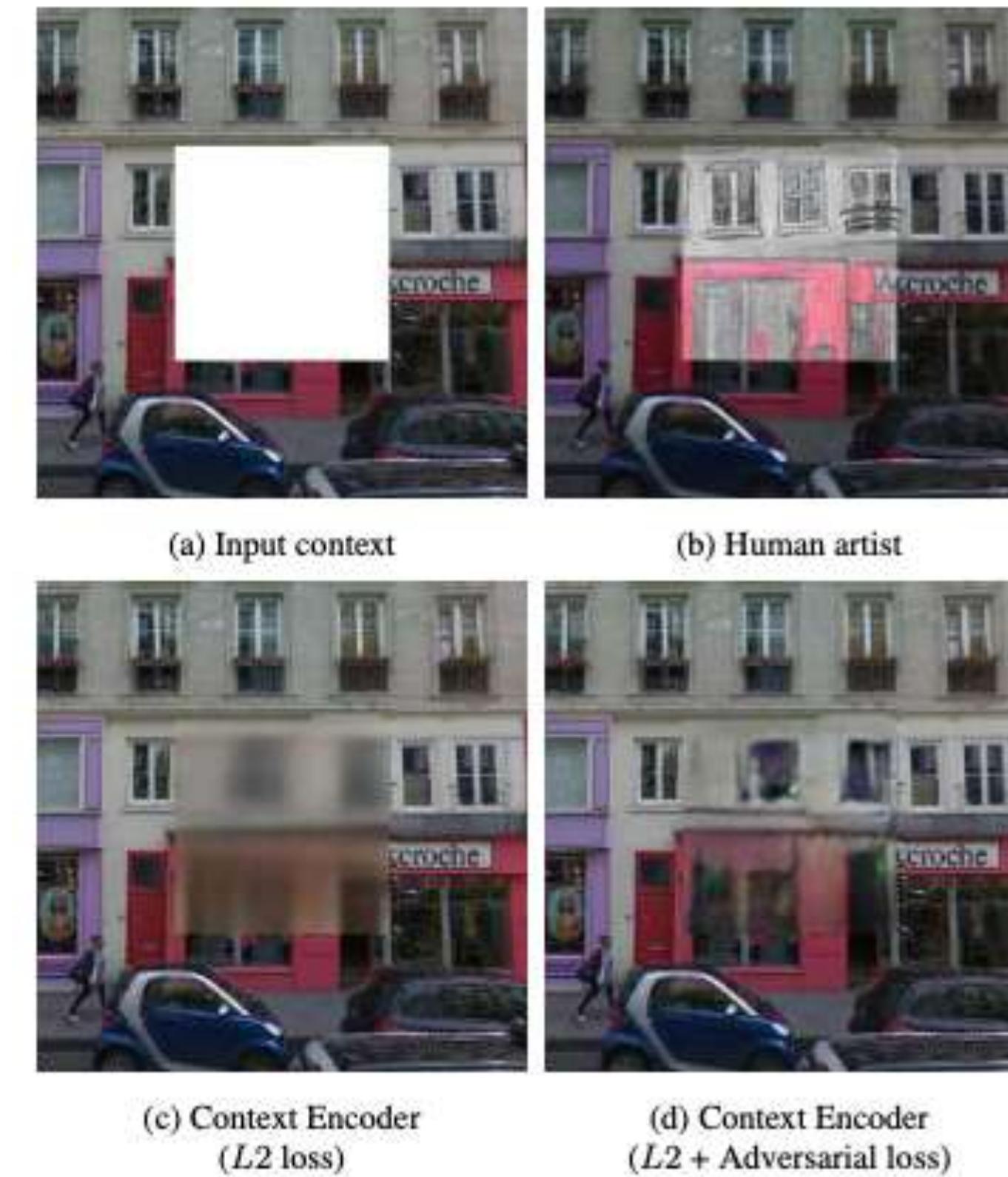
- Un autoencoder que recibe como entrada una muestra corrupta y se entrena para predecir los datos original no corrupto
- Aparte del extractor de características se tiene un “denoiser” como bono
- Brecha entre entrenamiento y uso: imagen de entrada en entrenamiento con ruido
- Demasiado fácil, sin necesidad de semántica: basta con características de bajo nivel



Generativos

Inpainting

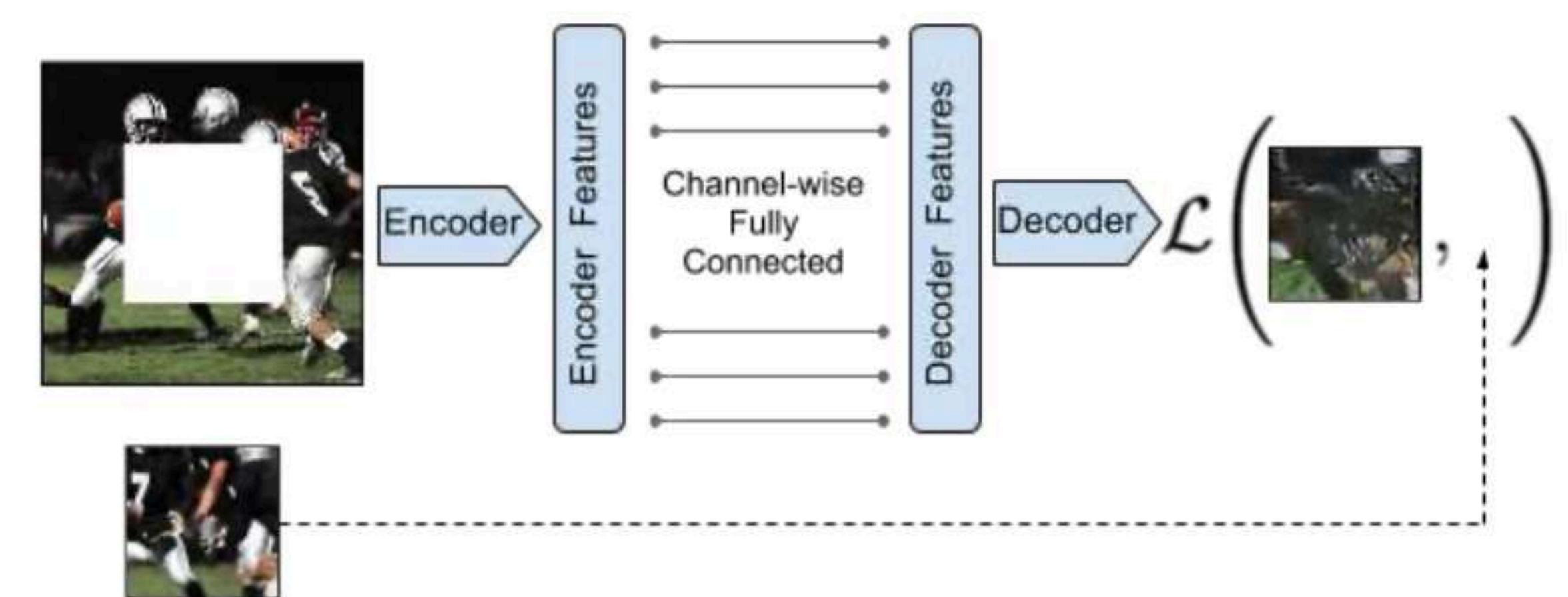
- Generar el contenido de una región arbitraria de la imagen en función de su entorno
- Modelo debe comprender el contenido de toda la imagen y elaborar una hipótesis plausible sobre la parte o partes que faltan



Generativos

Inpainting

- La arquitectura general es un simple codificador-decodificador
- El encoder y decoder se conectan con “channel-wise fully-connected layer” que permite a cada unidad del decodificador razonar sobre el contenido completo de la imagen
- Reconstruction loss: distancia L_2 o Adversarial loss

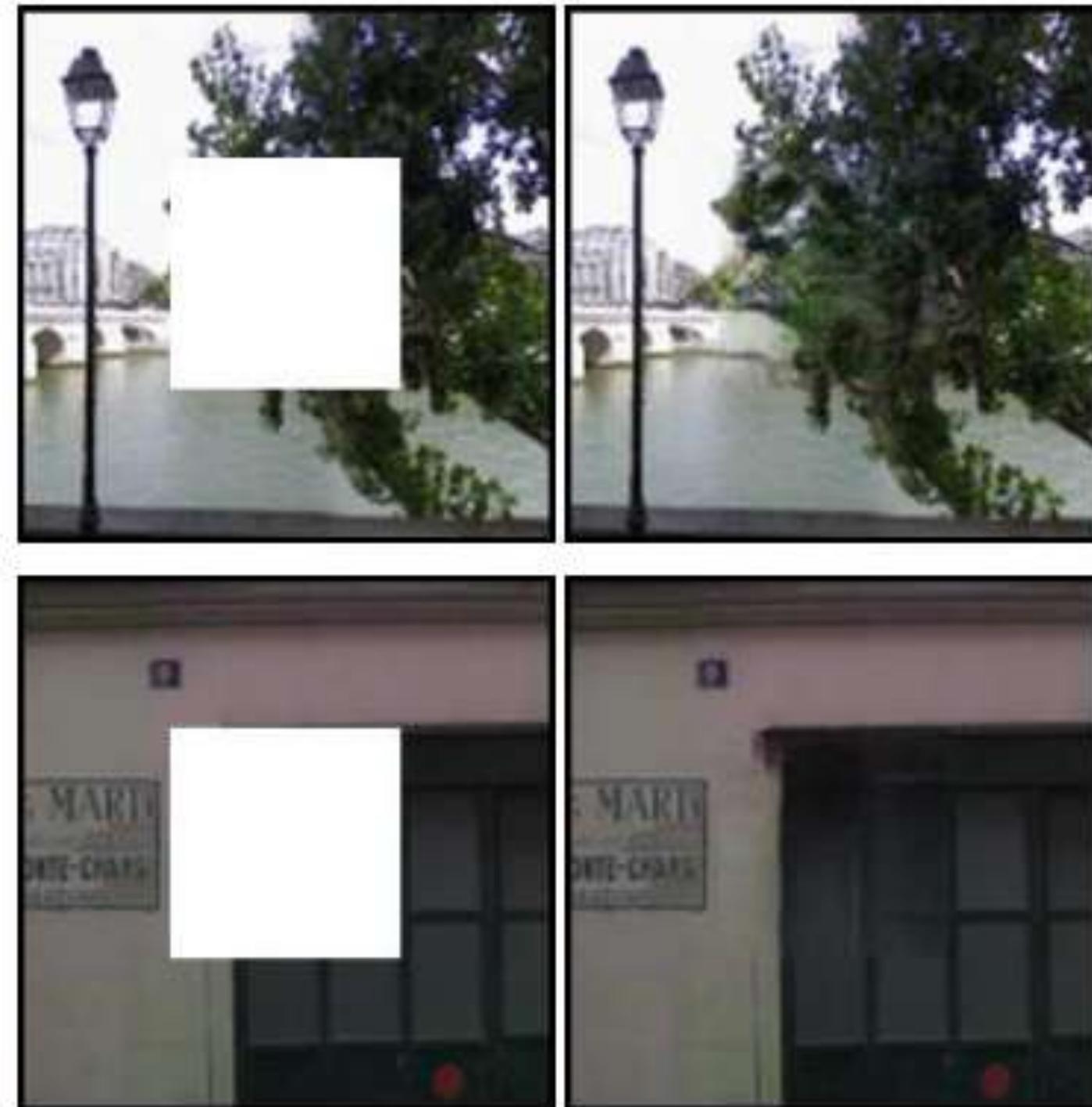


Función de costo
$$L_2(Y, \hat{Y}) = \|Y - \hat{Y}\|_2^2$$

Generativos

Inpainting

- Motiva a la conservación de información detallada
- Brecha entrenamiento-evaluación: no hay enmascaramiento en uso
- La reconstrucción es demasiado difícil y ambigua
- Se invierte mucho esfuerzo en detalles “inútiles” (Ejem: color exacto)



Generativos

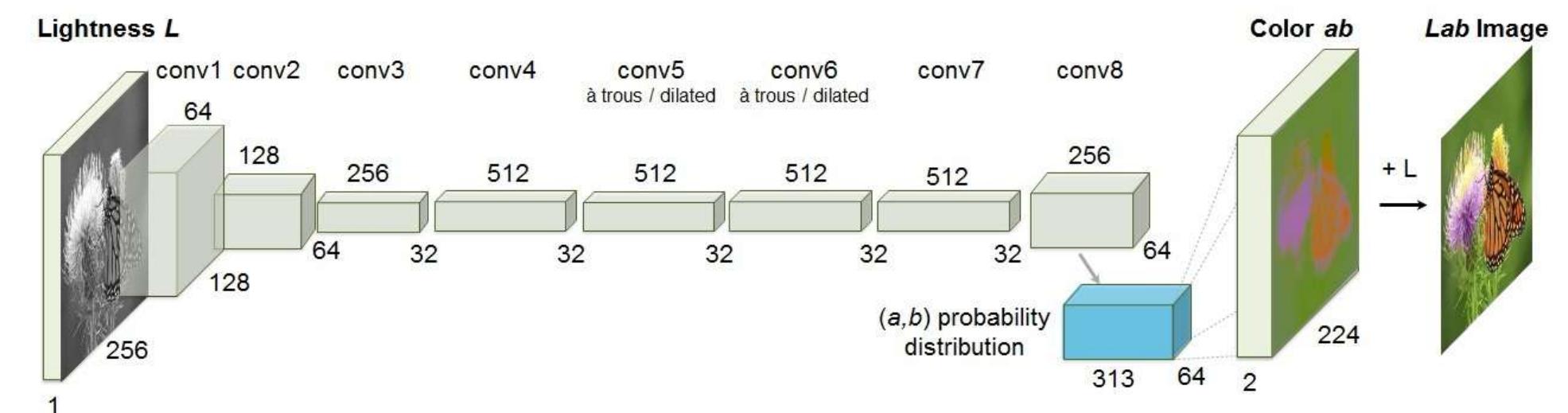
Colorization

- Dada una imagen en escala de grises como entrada, “alucinar” una versión en color plausible de la imagen
- Este problema está claramente infra-restringido (una manzana puede ser roja o verde)
- Mucha información se ha perdido, pero la semántica y textura pueden proveer suficientes pistas
- **El objetivo no es producir recuperar el color actual, sino, producir colores probables**

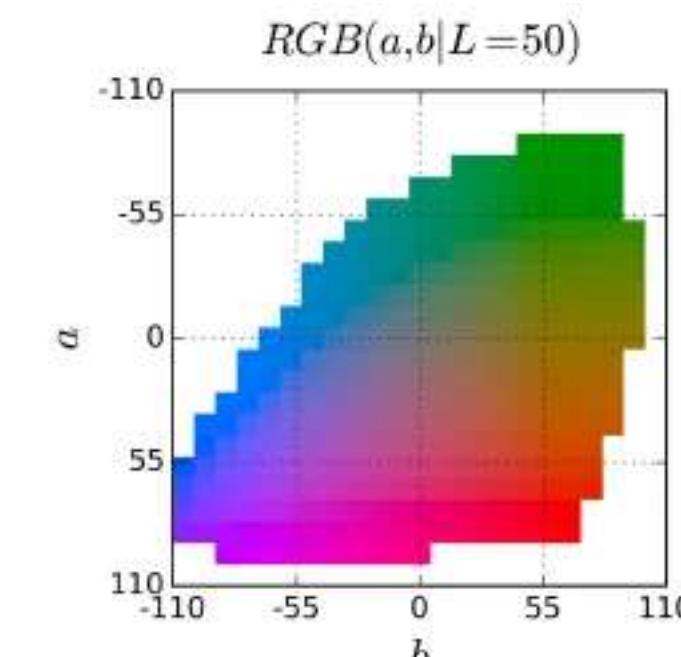


Generativos Colorization

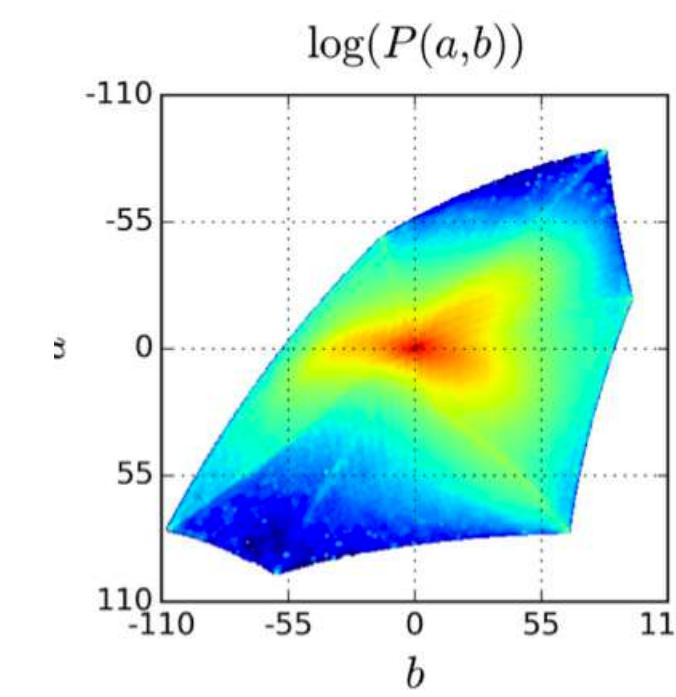
- Si se usa L_2 optimizaría el promedio de posibles colores y resulta en colores grises
- Tratar el problema como una clasificación y discretizar el espacio de color (313 clases pares de a,b)
- Se usa “weighted cross entropy loss” para penalizar colores comunes y promover colores raros (pre-calculado)
- Las etiquetas corresponden a un soft-encoding de los 5 colores más cercanos



Arquitectura de CNN



Espacio de color LAB cuantizado

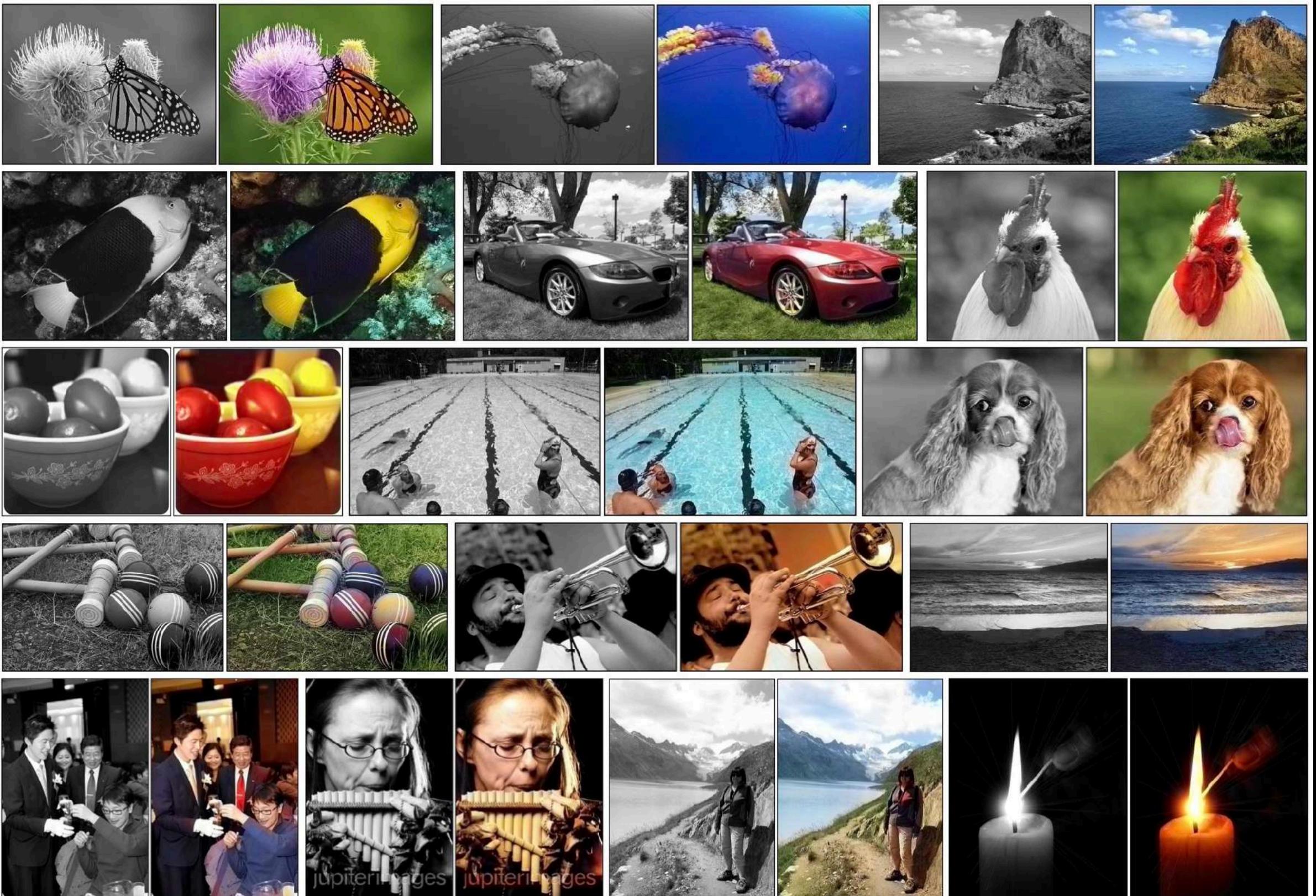


Probabilidad de colores (Empirica - ImageNet)

Generativos

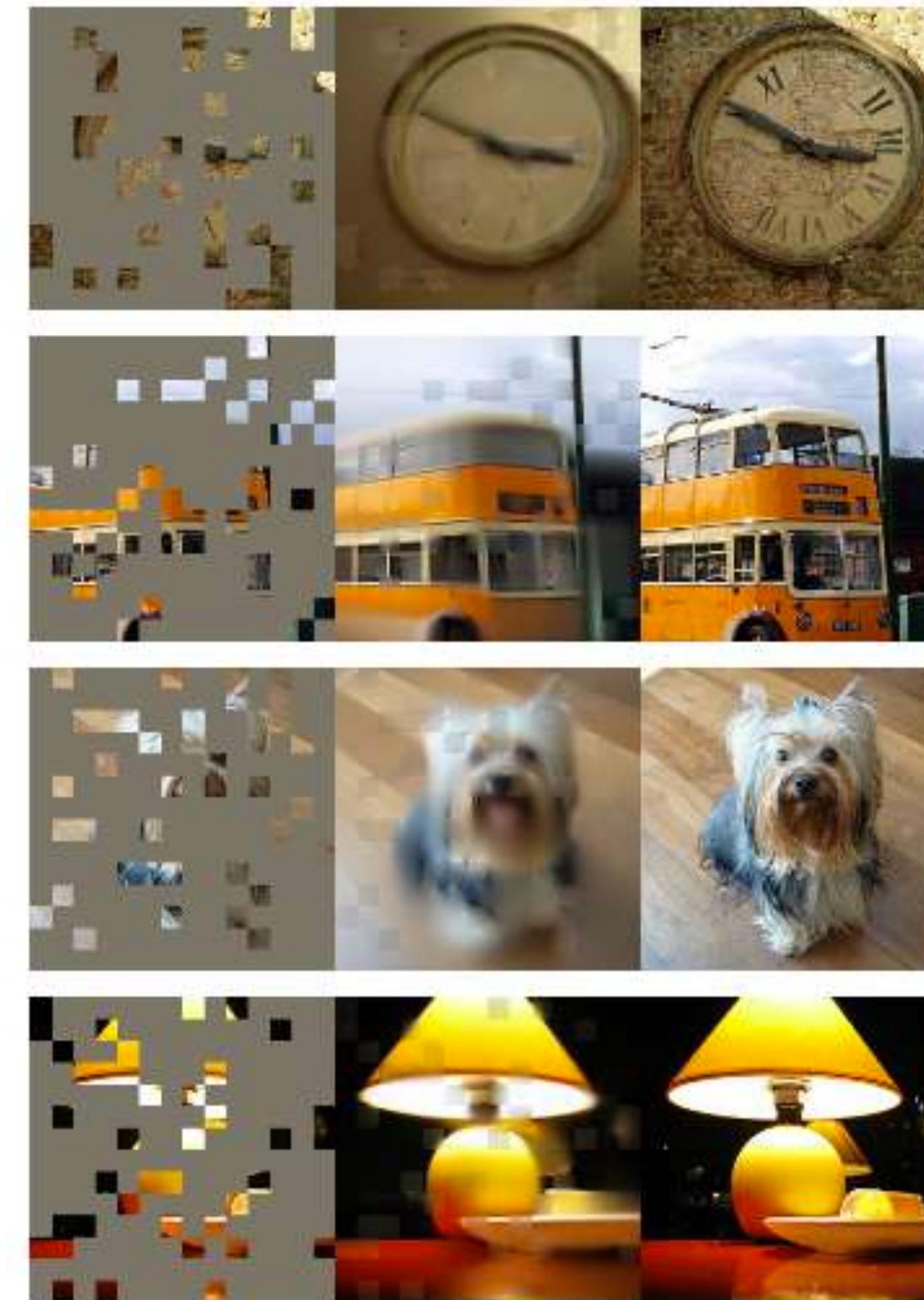
Colorization

- Requiere la conservación de información detallada en el espacio embebido
- La reconstrucción es demasiado difícil y ambigua
- Se gasta mucho esfuerzo en detalles “inútiles” (Ejem: color exacto)
- Brecha entrenamiento-evaluación: durante entrenamiento imágenes en grises



Generativos Masked Autoencoders

- Inspirado en “Masked Language Modeling” usado en BERT
- Aplica máscaras aleatorias a los datos de entrada y el modelo reconstruye las partes faltantes



2021 - Masked Autoencoders Are Scalable Vision Learners

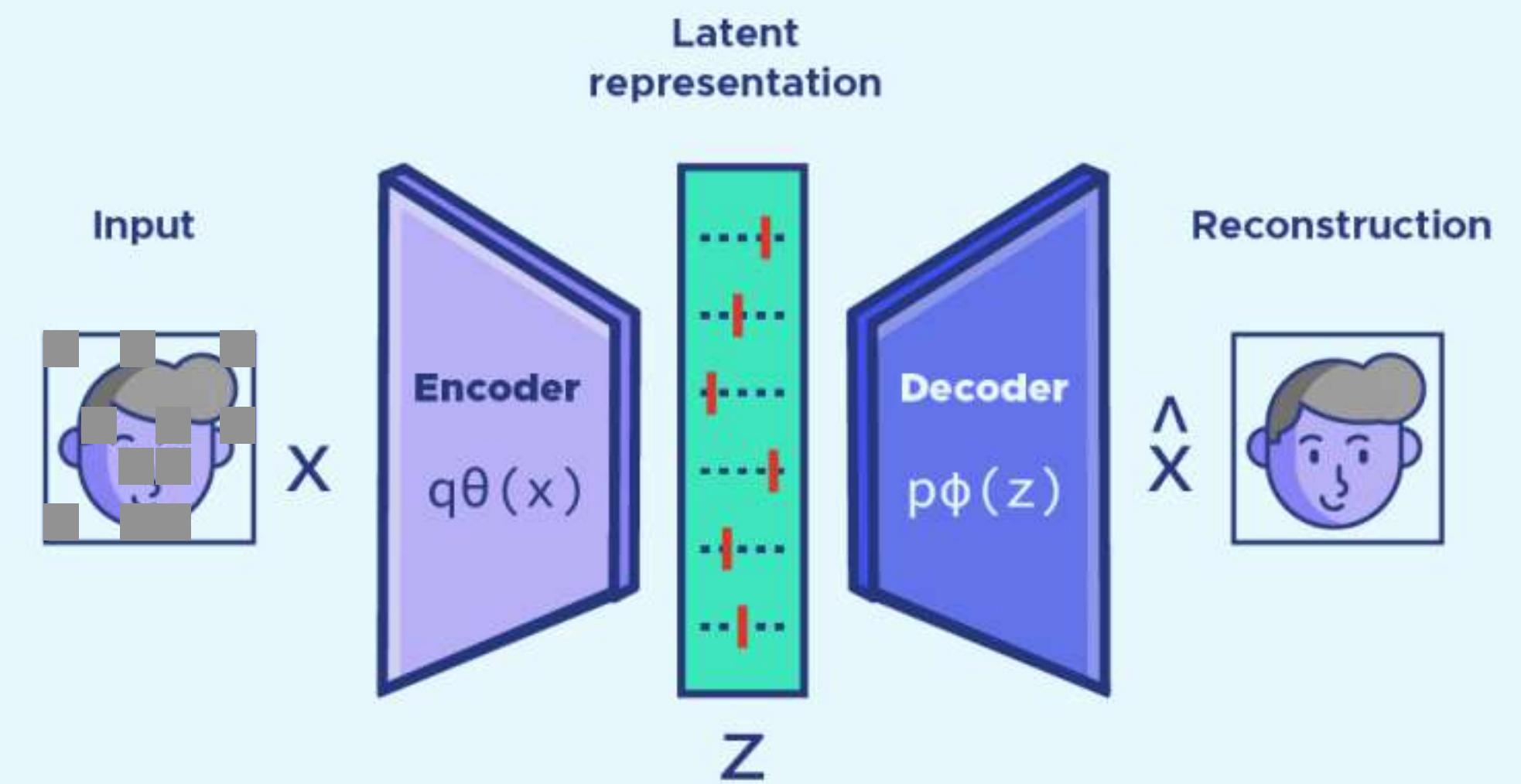
MAE

Arquitectura

Encoder: que mapea una señal observada a una representación latente

Decoder: reconstruye la señal original desde la representación latente

Diferencia: diseño asimétrico, el encoder opera solamente con observaciones parciales y el decoder reconstruye toda la observación



MAE

Arquitectura

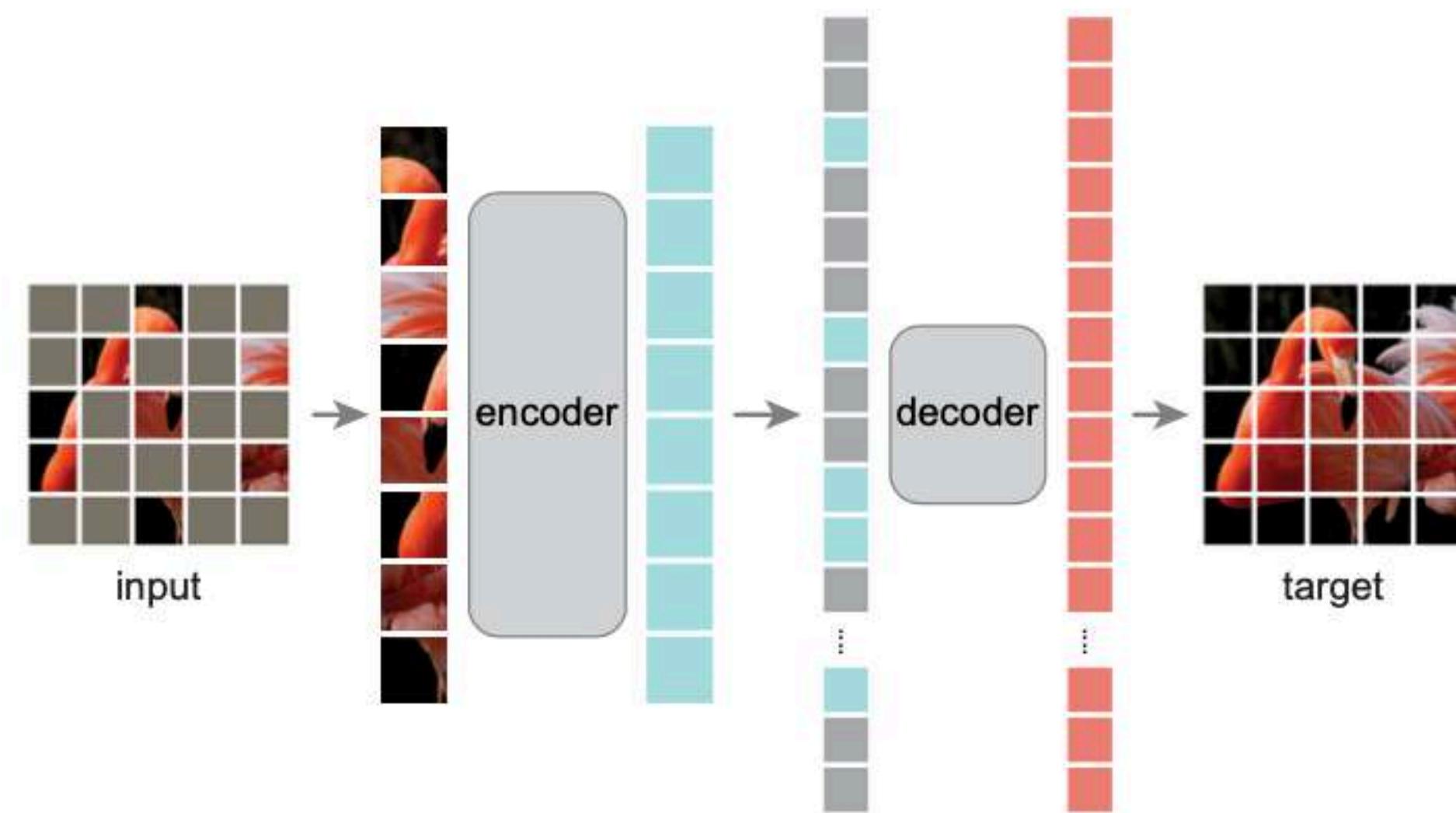
Masking: divide una imagen en parches regulares no traslapados. Se muestrea parches aleatorios sin reemplazo

MAE encoder: ViT pero solo aplicado a los parches visibles

MAE decoder: ViT, la entrada es el conjunto entero de tokens:

- Parches visibles codificados
- Mask tokens: vectores aprendidos que indican la presencia de un parche que debe ser reconstruido

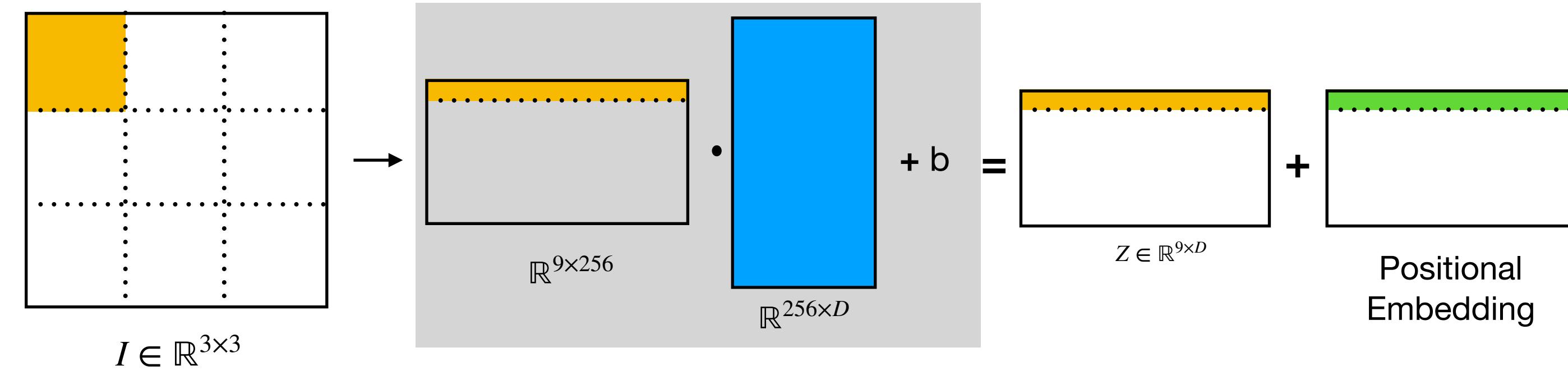
El decoder es solo usado durante la etapa de entrenamiento para la tarea pretexto de reconstrucción



MAE

Implementación

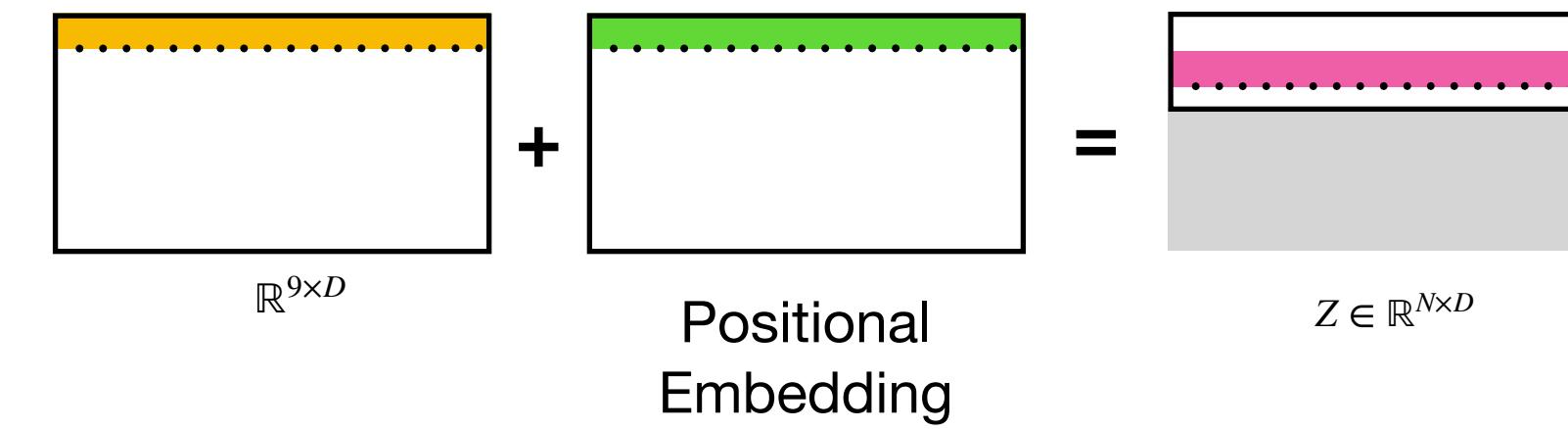
Embebido ocurre igual que en ViT



Image

Patch Embedding

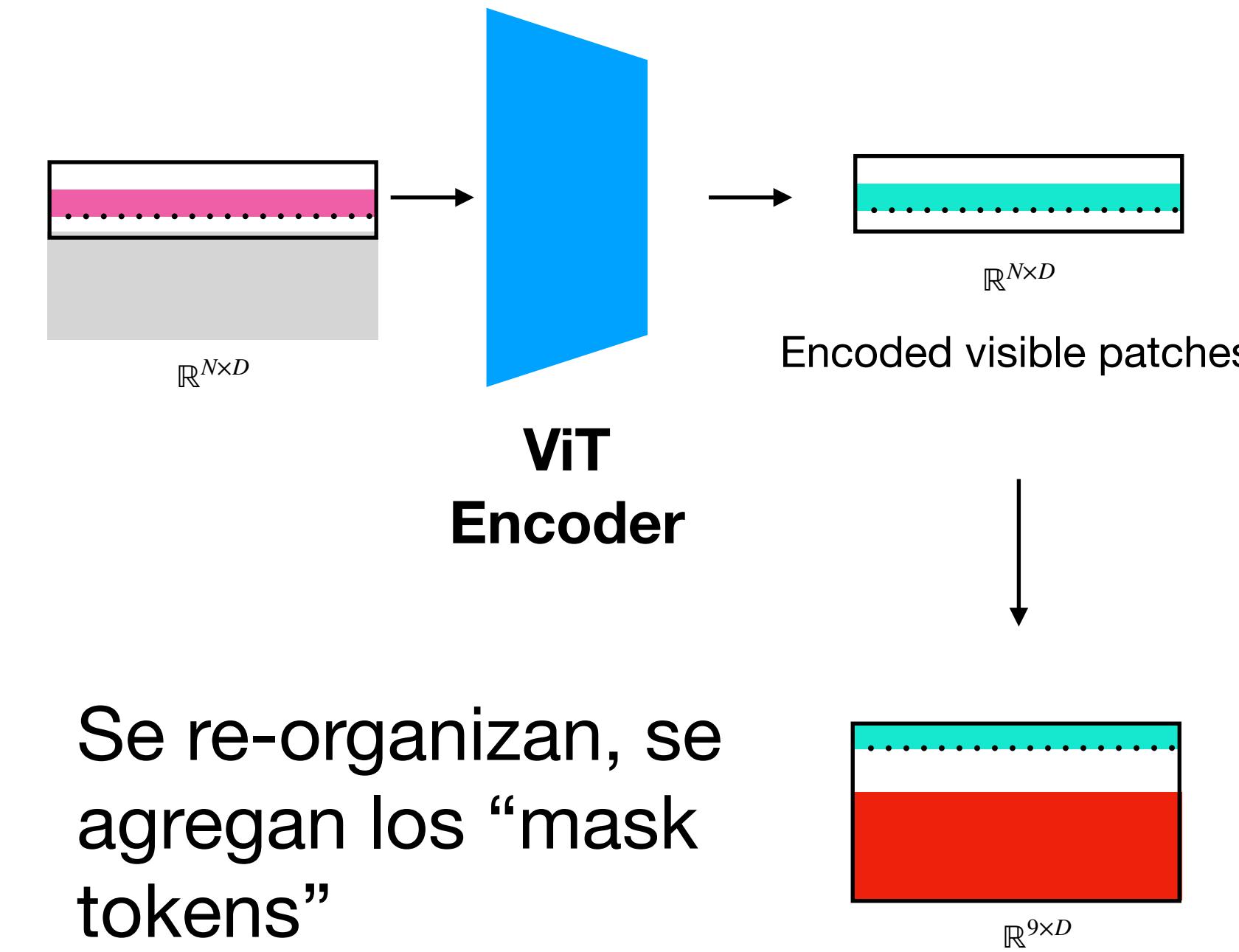
Se reorganizan aleatoriamente y se remueven los ocultos



MAE

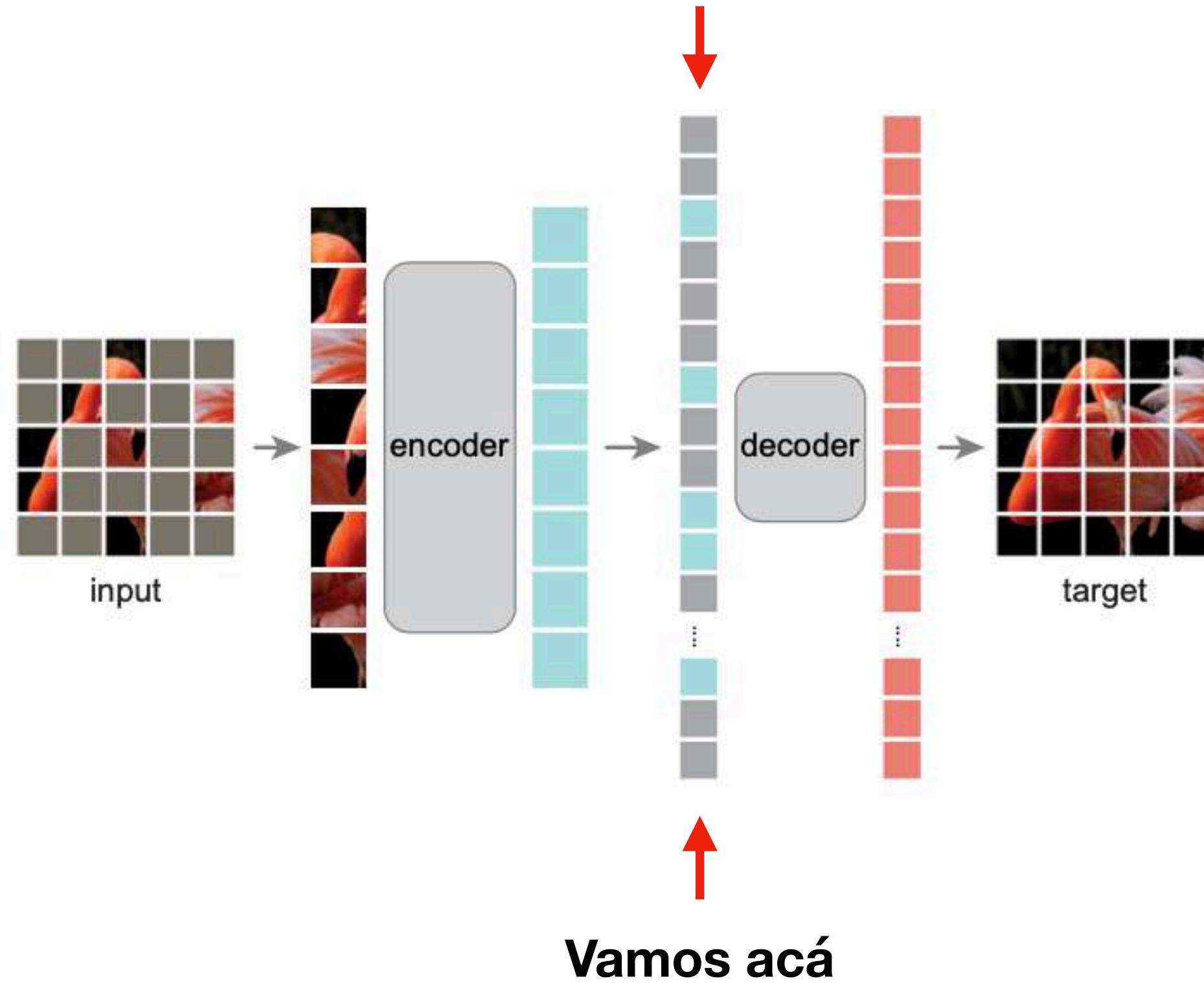
Implementación

N embebidos visibles
van al encoder



MAE

Implementación

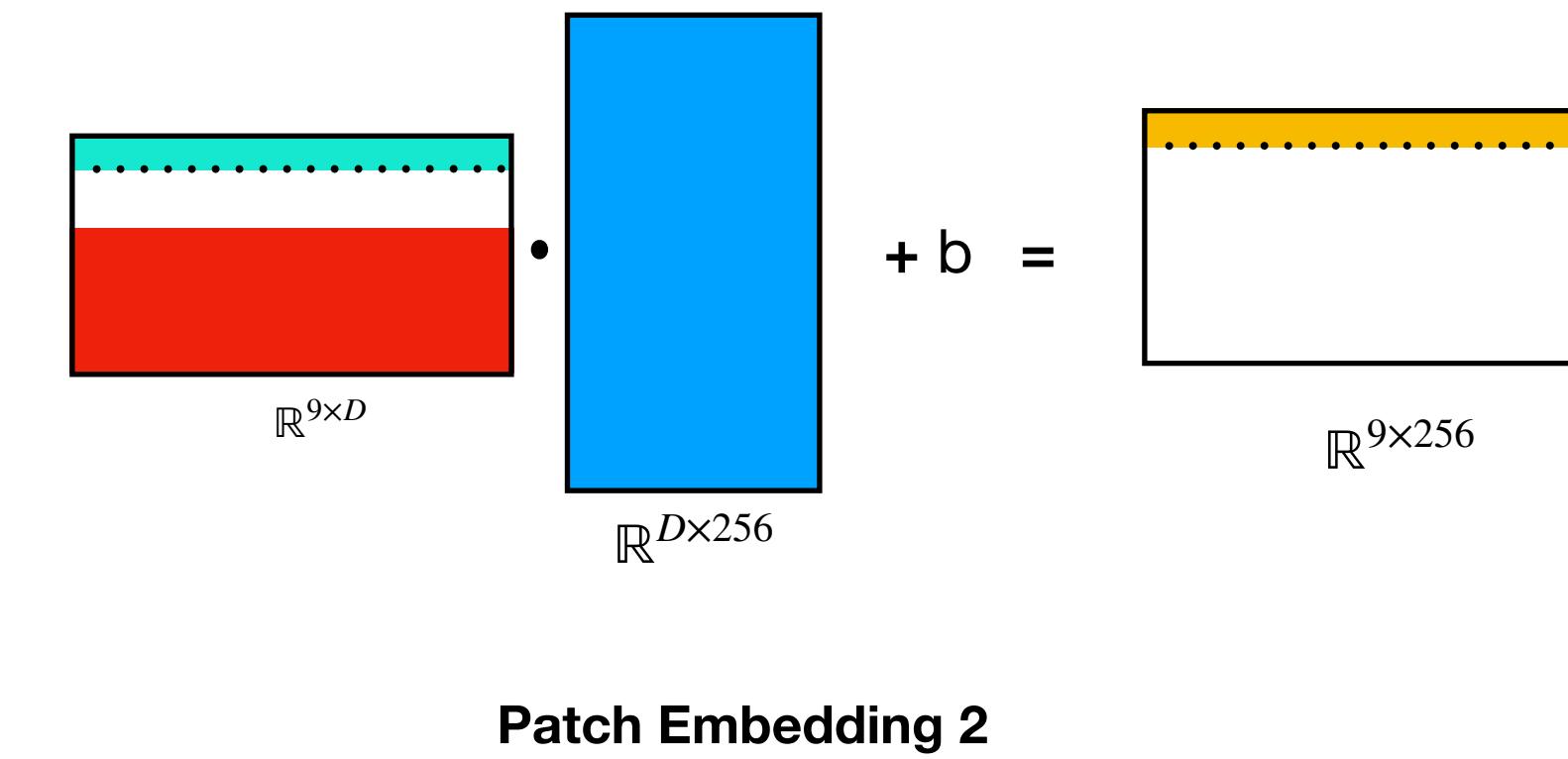


MAE

Implementación

Como el Decoder es un ViT se realiza primero un embebido de parches.

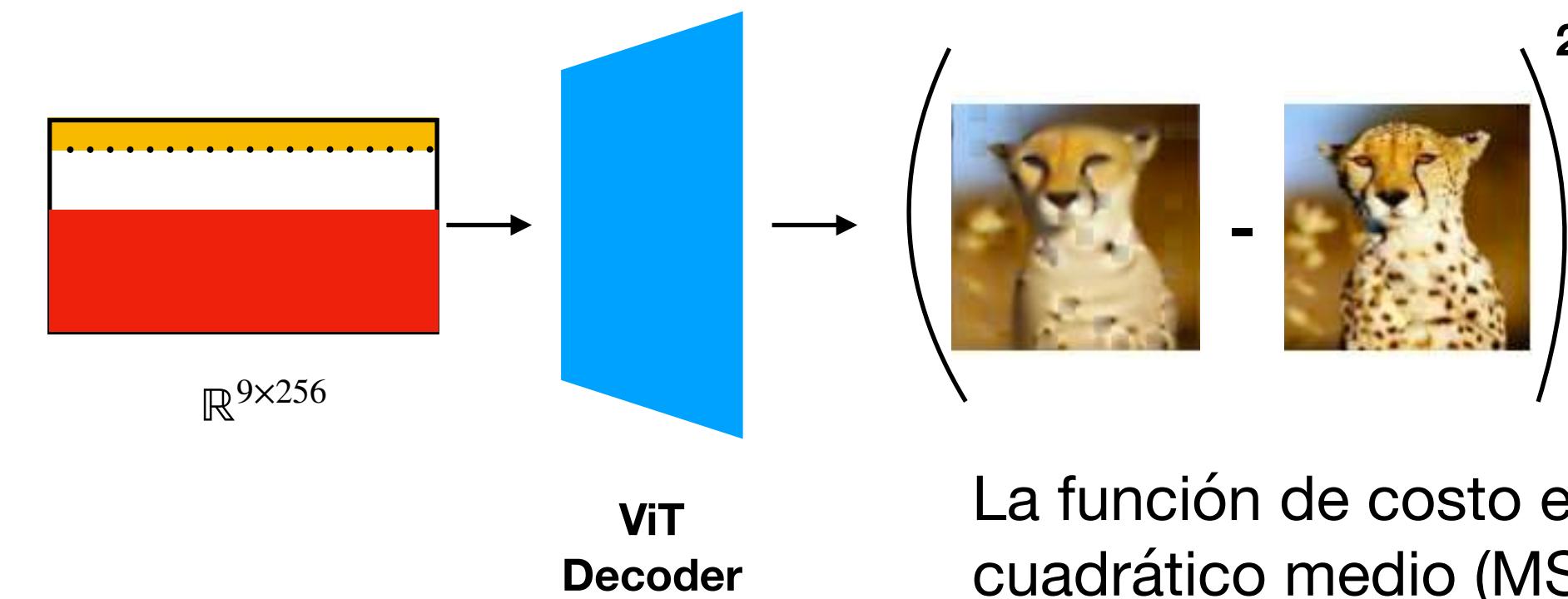
Se encarga de la expansión dimensional al tamaño inicial de los patches



MAE

Implementación

La salida del decoder se organiza para formar una imagen reconstruida



La función de costo es el error cuadrático medio (MSE) entre las imágenes reconstruida y original

Agenda



- Modelos fundacionales
- Self-Supervised Learning
 - Predictivos - Pretext tasks
 - Contrastivos (SimSLR)
 - Generativos (MAE)
- **Discusión**
- Taller

Retos

Pre-entrenamiento de grandes a gran escala



- Al ser conjuntos de datos masivos pueden contener imágenes etiquetadas incorrectas o irrelevantes
- La captación de datos pueden reforzar inadvertidamente sesgos sociales que influencien las respuestas del modelo
- Enormes recursos computacionales, accesibles solo a organizaciones bien financiadas
- Entrenamiento inestable que requiere cuidado extra en la optimización de hiperparámetros
- Preocupante consumo de energía y huella de carbono

SSL vs TL

Self-Supervised Learning vs Transfer Learning

En general:

- TL funciona mejor con dominios cuando los dominios y tareas son similares
- SSL es más robusto frente a diferentes tareas, dominios y desequilibrio de clases

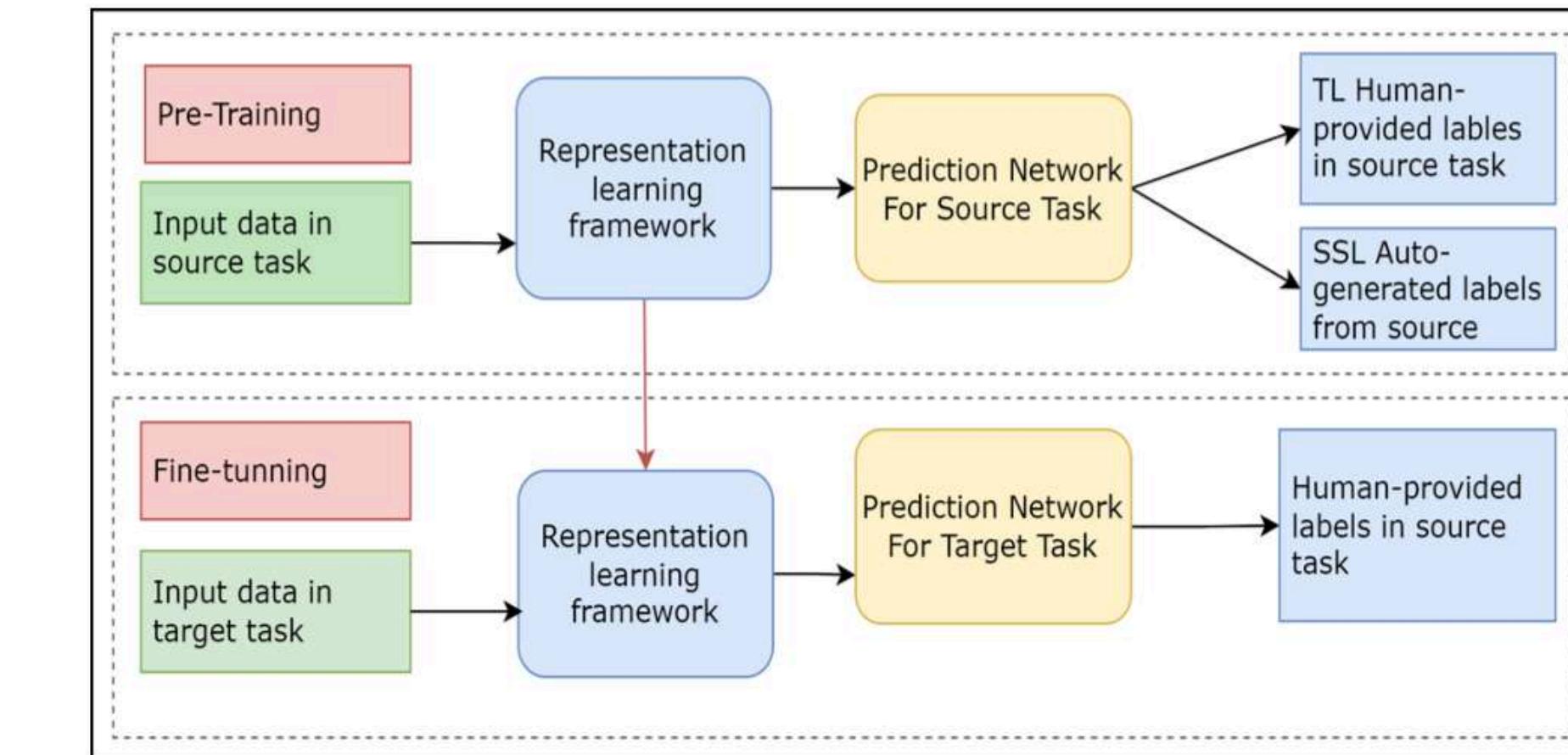


Fig. 3: TL and SSL workflow.

PFM vs CNN

Aspecto	Classical CNNs	Vision Foundational Models
Arquitectura	Basado en capas convolucionales	Basado en bloques multi-head self-attention
Entrenamiento	Aprendizaje supervisado (e.g., ImageNet)	Aprendizaje auto-supervisado (pretext tasks)
Escalabilidad	La complejidad crece con la profundidad y el tamaño de entrada; escalabilidad limitada	Diseñados para escalabilidad, manejando eficientemente datasets grandes
Versatilidad	Especializado para tareas basadas en imágenes; aprendizaje por transferencia	General-purpose y multimodal, capaz de few-shot o zero-shot learning
Rendimiento	Efectivo para tareas bien definidas con datos etiquetados	Robustos en escenarios reales con cambios de dominio.
Computo	Más ligeros computacionalmente, prácticos para dispositivos con recursos limitados.	Requieren altos recursos computacionales para pre-entrenamiento

Agenda



- CNN Inductive Bias
- Vision Image Transformers (ViT)
- Modelos fundamentales
- Self-Supervised Learning
- Discusión
- **Taller**

Taller

- Implementar un modelo CNN o ViT.
Puede ser un modelo de alguna librería
(Keras) o una implementación “vanilla”
de una CNN
- Inventar una tarea de pretexto
- Entrenar en ImageNet



Referencias

- 2020 - Contrastive Representation Learning A Framework and Review
- 2023 - Foundational Models Defining a New Era in Vision A Survey and Outlook
- 2021 - Masked Autoencoders Are Scalable Vision Learners
- 2017 - Attention is all you need
- 2021 - An Image Is Worth 16X16 Words Transformers For Image Recognition At Scale
- 2020 - A Simple Framework for Contrastive Learning of Visual Representations
- 2024 - A Survey of the Self-Supervised Learning Mechanisms for Vision Transformers
- 2016 - Unsupervised Visual Representation Learning by Context Prediction
- 2017 - Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles
- 2018 - Unsupervised representation learning by predicting image rotations
- 2008 - Extracting and Composing Robust Features with Denoising Autoencoders
- 2016 - Context Encoders: Feature Learning by Inpainting