Self-supervised learning
SSL

semiDPS 2025

Sesión 1

"Organización" del seminario

- ***¿Qué es SSL, para qué y cómo?***
  Survey (primeras dos sesiones)
  Cookbook ¿?

- ***Revisión*** de:
  - artículos de interés
  - trabajos del DPS con este enfoque

Guías para el seminario

- A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends
- A Cookbook of Self-Supervised Learning

Nota: nomenclatura variada para diferente bibliografía

¿Qué es SSL, para qué y cómo?

1. Introducción
    a. Por qué SSL
    b. Esquema de SSL

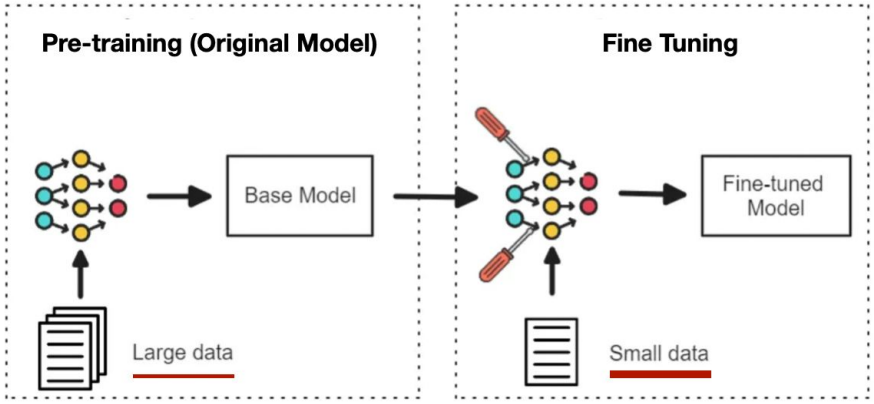
2. Algoritmos / pretext-tasks
    a. Context-based
    b. Contrast Learning
    c. Generative / Masked Image Modeling (MIM)

Introducción

DL supervised da buenos resultados para CV, NLP. Procedimientos de transfer learning sobre modelos entrenados en **datasets etiquetados de grandes volúmenes:**

a. partir de un pre trained model → buena inicialización y "asegurarse" convergencia

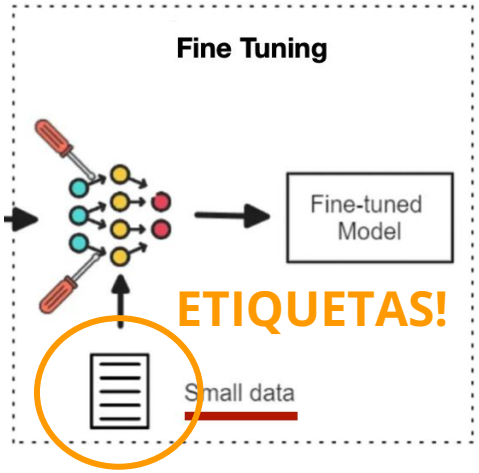
b. features del modelo son buenos discriminadores para que se transfieran a otro **objetivo** final o ***"downstream task"***



Introducción

Problema: *necesidad de un conjunto de datos totalmente etiquetados.*
Estos suelen ser:

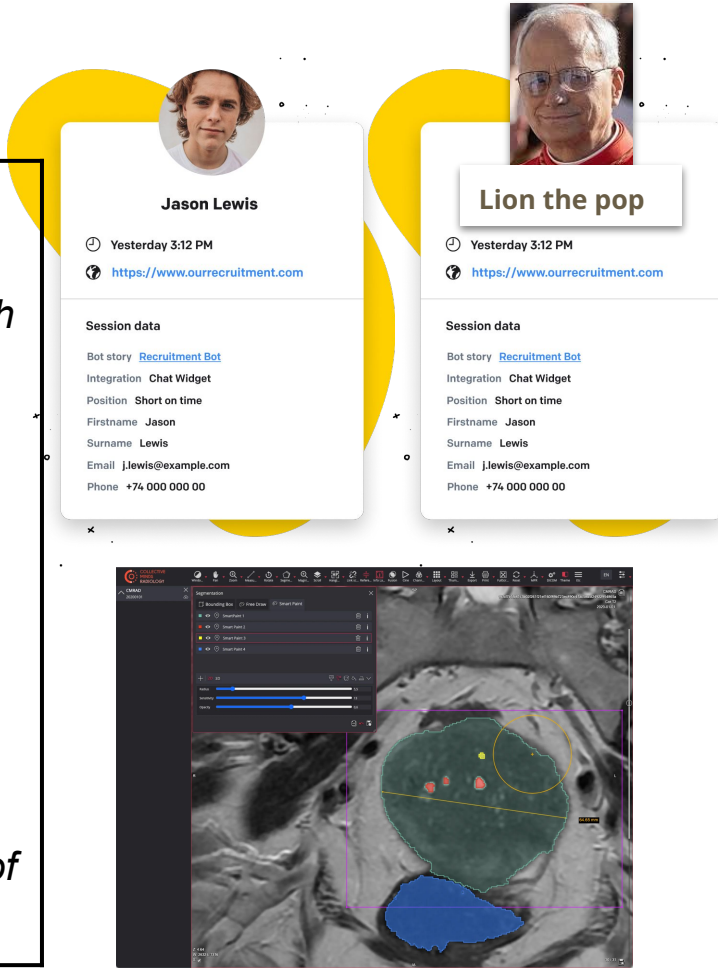
- costosos / poco acceso
- necesidad de expertos
- consumen mucho tiempo en adquirir
- a veces erróneos (correlaciones espúreas)



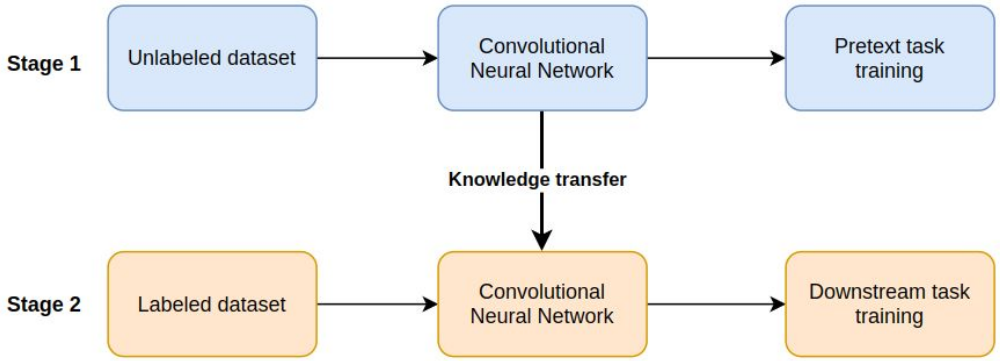
Introducción

*"… consider the **analysis of web user profiles**, where a substantial amount of data can be readily collected. However, the labeling of non-profitable or profitable users necessitates thorough scrutiny, judgment, and sometimes even time intensive tracing tasks performed by experienced human assessors, resulting in significant expenses.*

*Another instance pertains to the **medical field**, where unlabeled examples can be easily obtained through routine medical examinations. Nevertheless, assigning diagnoses individually to such a large number of cases places a substantial burden on medical experts. For example, in the case of breast cancer diagnosis, radiologists must label each focus in a vast collection of easily attainable, high-resolution mammograms."*



Introducción



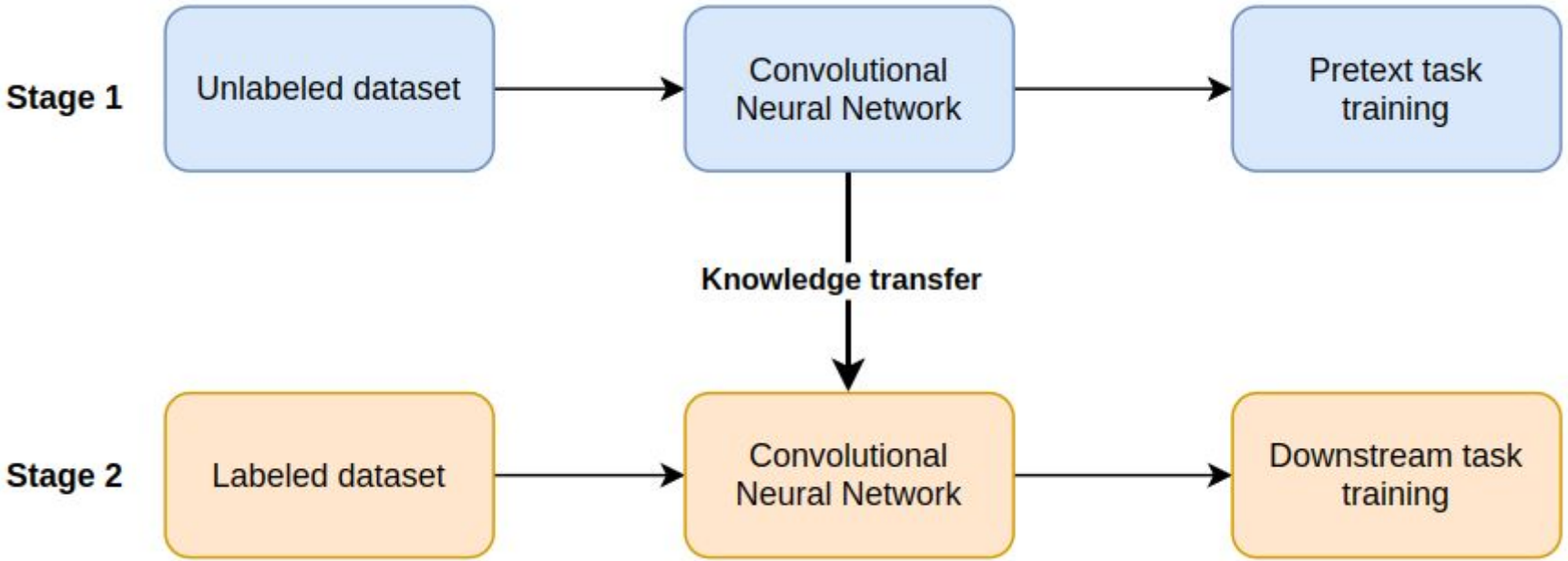
Esquema SSL

Para mitigar estas limitaciones, SSL se enfoca en aprender **características discriminativas** sobre grandes cantidades de datos no etiquetados de la siguiente manera:

- ***Pretext task***: generar pseudo-etiquetas según el tipo de dato, *SIN NECESIDAD DE ETIQUETAS*. Se termina con un modelo pre-entrenado para un problema diferente al objetivo end-to-end.

- ***Downstream task***: transferir el modelo aprendido en la etapa anterior para la tarea específica *CON POCOS DATOS ETIQUETADOS.*

Introducción



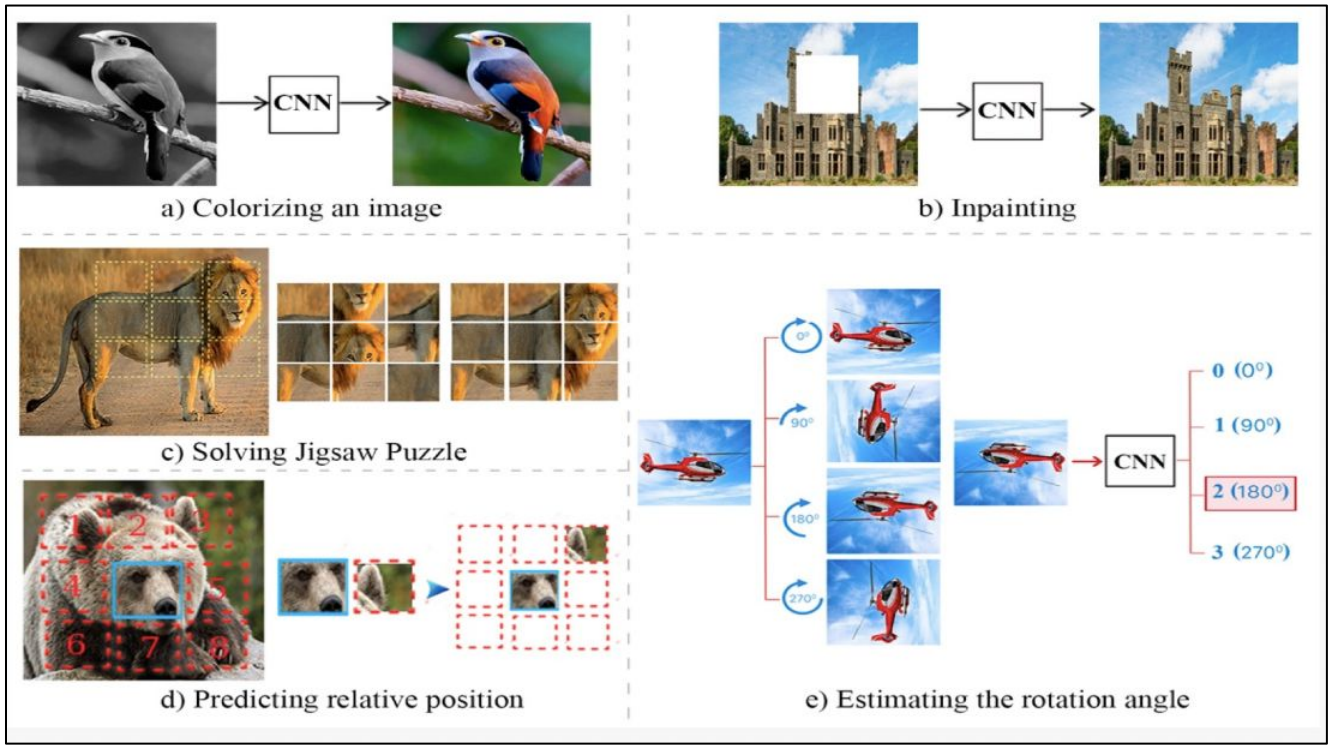
Introducción

En SSL, el pretext-task **genera etiquetas** que son ***coherentes*** con los datos de entrada.

Debe ser capaz de mostrar la relación ante diferentes **vistas** de la entrada. Así los modelos extraen buenas características de los datos al momento de resolver estas tareas.

DISCLAIMER: de ahora en más "*IMAGEN*" = *DATO GENÉRICO*

Introducción



Ejemplos de pretext-tasks

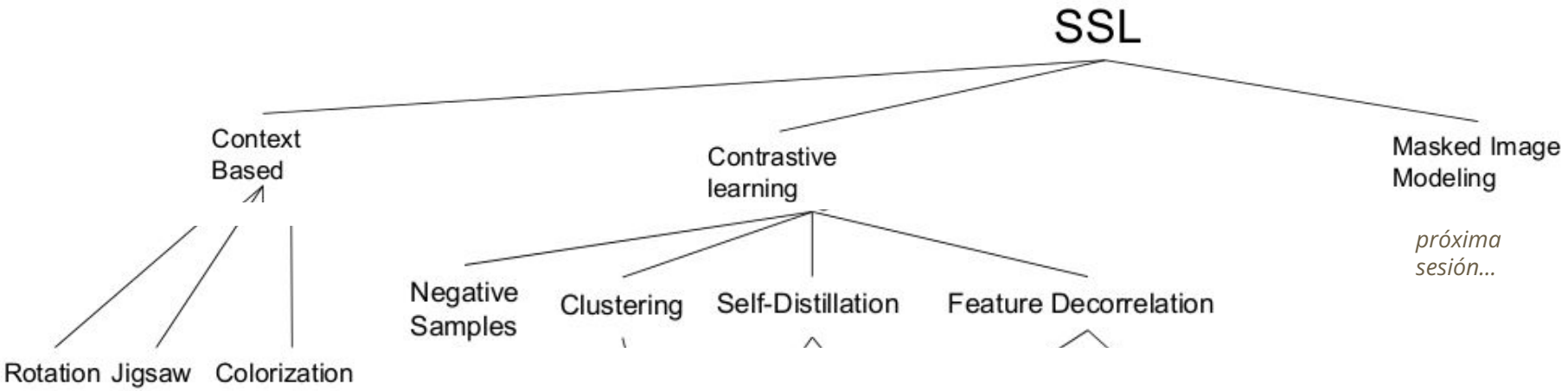
Introducción

El pretext task es un concepto **fundamental para SSL** ya que suele referirse a una tarea a resolver, que **no resuelve el objetivo principal de aplicación**.

En cambio, sí genera **modelos (robustos)** pre-entrenados **sin necesidad de etiquetas**. Para ello, las redes se entrenan optimizando funciones de pérdida (*loss*) asociadas a este tipo de tareas. Los distintos tipos de tareas pueden llegar a ser:

- ***Context-based methods***
- ***Contrastive Learning (CL)***
- ***Generative algorithms o Masked Image Modeling (MIM)***

Introducción

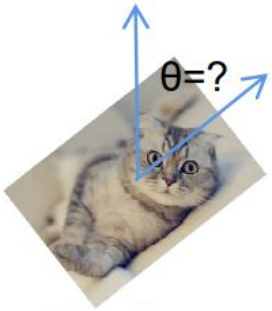


próxima sesión...

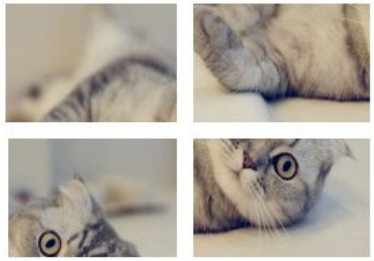
Algoritmos: Context-based

Buscan resolver problemas de contexto sobre el mismo dato donde se aplica una transformación y la red predice el dato mismo o cómo fue dicha transformación:

- encontrar ángulo rotación de imagen
- resolver orden de un “puzzle/jigsaw”
- encontrar canales croma *ab* dado el canal de luminancia *L* de una imagen



Rotation

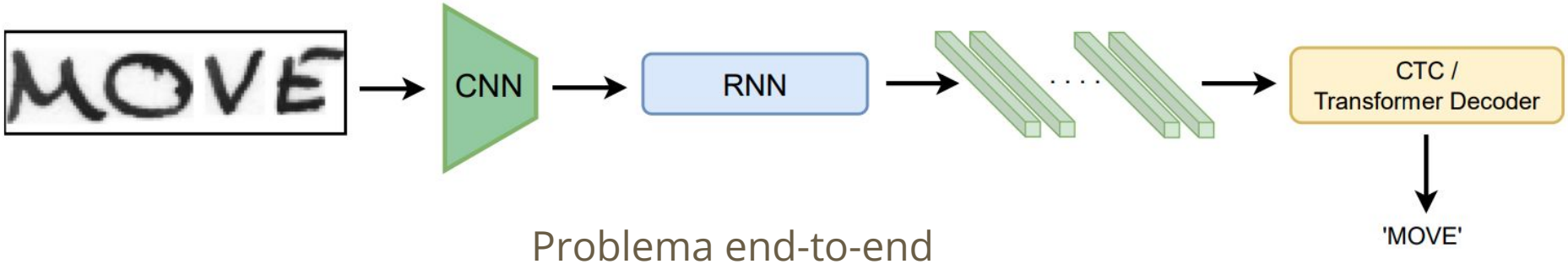

Jigsaw


Colorization

Discutir cómo serían estos pretextos

Algoritmos: Context-based

Ejemplo completo para aplicación en reconocimiento de texto manuscrito

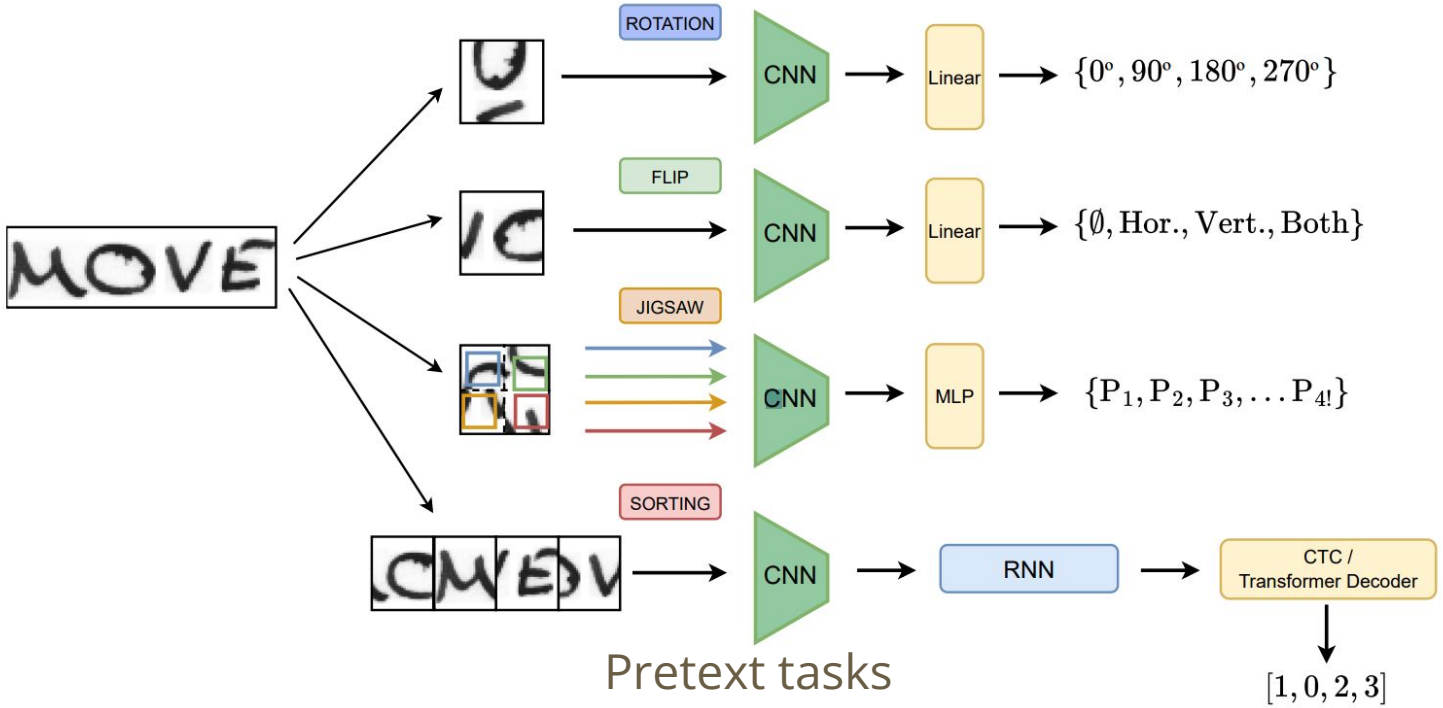


Problema end-to-end

Se emplea un pretext-task sobre la red 

Algoritmos: Context-based

Ejemplo completo para aplicación en reconocimiento de texto manuscrito



Pretext tasks

Algoritmos: Context-based

Resultados sobre todos los datasets **(enteros)** no muy prometedores.

Métrica: WAcc

Method	CTC		
	IAM	CVL	RIMES
Baselines			
SeqCLR†	39.7	66.7	63.8
ChaCo†	**70.0**	**75.6**	79.4
CMT-Co†	53.1	72.1	66.0
Text-DIAE‡	-	-	-
Geometric			
Rotation‡	65.2	54.9	**85.0**
Flip‡	65.9	55.4	82.0
Puzzle			
Jigsaw‡	36.4	18.1	53.9
Sorting‡	55.0	26.4	66.1

Algoritmos: Context-based

Metrica Word Accuracy sobre diferentes datasets (5, 10 y 100 %) en downstream task.

Method	IAM			CVL			RIMES		
	5	10	100	5	10	100	5	10	100
Baselines									
SeqCLR	31.2	44.9	76.7	66.0	71.0	77.0	61.8	71.9	90.1
ChaCo	44.3	53.1	79.8	65.6	70.3	78.1	61.3	73.0	89.2
CMT-Co	47.7	56.0	**80.1**	**71.9**	**74.5**	78.2	67.1	75.4	90.1
PerSec-CNN	-	-	77.9	-	-	78.1	-	-	-
Geometric									
Rotation	60.2	66.3	79.0	39.0	52.5	85.3	79.7	**84.4**	**91.4**
Flip	**61.7**	**67.4**	79.1	40.6	54.6	**85.5**	**79.8**	84.2	91.1
Puzzle									
Jigsaw	54.5	64.0	77.7	28.7	42.1	83.8	77.6	81.0	90.8
Sorting	56.4	64.7	79.0	31.8	34.7	78.5	74.1	71.5	90.0

pretextos

Algoritmos: Context-based

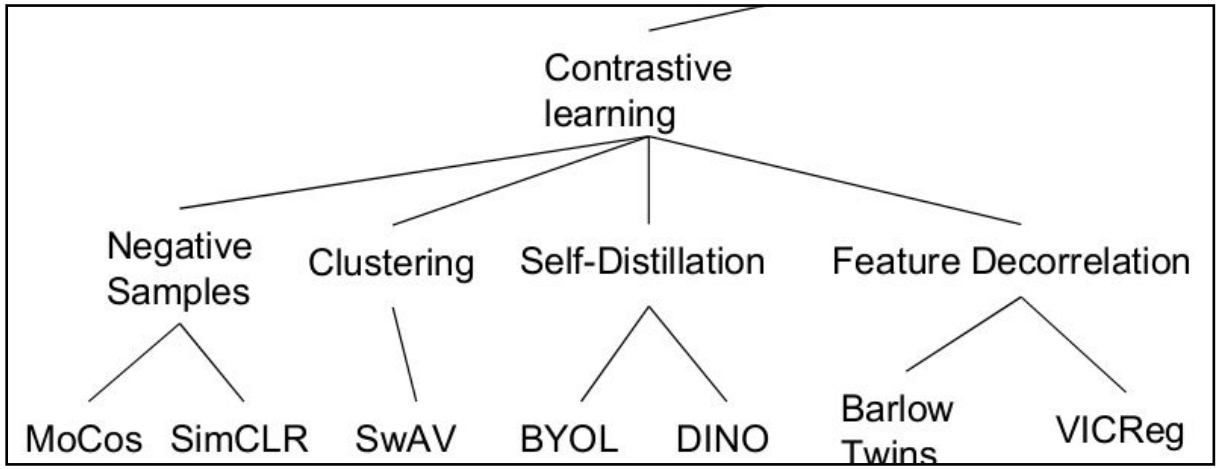
(a) Rotation (b) Flip

(c) Jigsaw (d) Sorting

Figure 4: Heat map highlighting the regions targeted by the encoder pre-trained with the different spatial context-based SSL methods.

Contrastive Learning

Técnica que busca mantener consistencia positiva entre los datos que correspondan a lo mismo, a pesar de una manipulación.



Contrastive Learning - Negative example-based

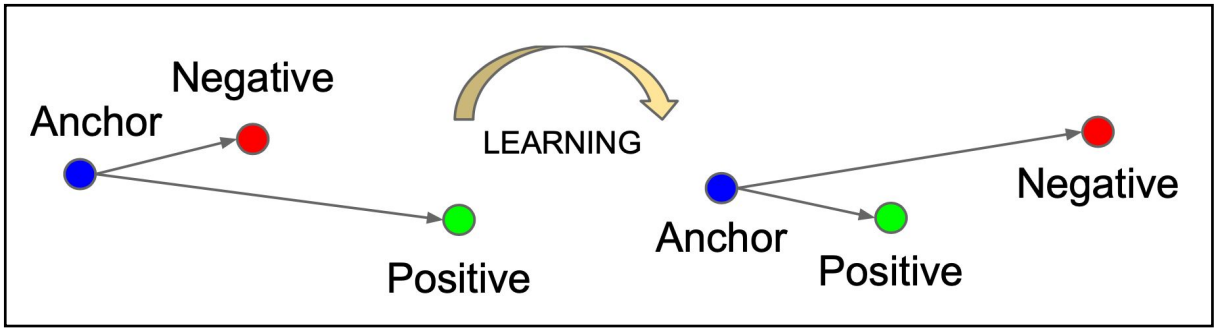
Utiliza discriminación de instancia.

- Dado un dato o instancia “ancla” se originan diferentes vistas, consideradas como ejemplos positivos.
- Se obtienen ejemplos negativos a partir de otras instancias.

En el espacio de latencia:

- promueve proximidad entre las muestras positivas y el ancla
- maximiza separación entre muestras negativas y el ancla

La definición de muestras positivas y negativas van a depender del tipo de dato y los requerimientos que se buscan preservar. Aquí se usa la loss InfoNCE/NT-XEnt:



Contrastive Learning - Negative example - SimCLR

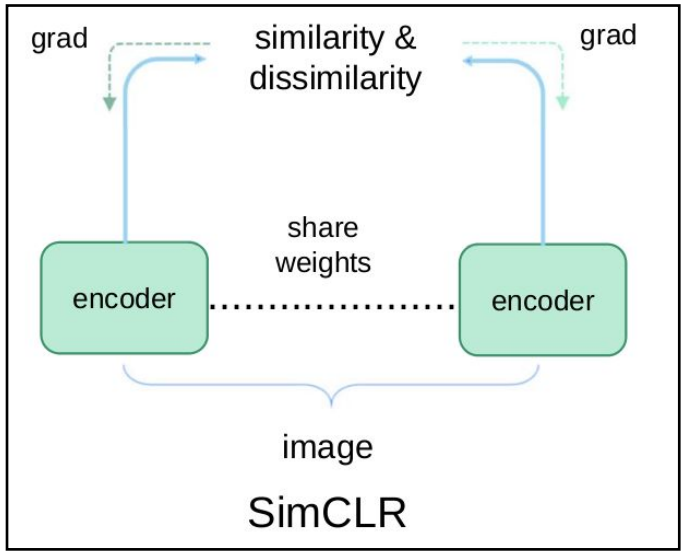
Utiliza discriminación de instancia.

- Dado un dato o instancia “ancla” se originan diferentes vistas, consideradas como ejemplos positivos.
- Se obtienen ejemplos negativos a partir de otras instancias.

En el espacio de latencia:

- promueve proximidad entre las muestras positivas y el ancla
- maximiza separación entre muestras negativas y el ancla

La definición de muestras positivas y negativas
van a depender del tipo de dato y los requerimientos
que se buscan preservar. Aquí se usa la loss InfoNCE/NT-XEnt:



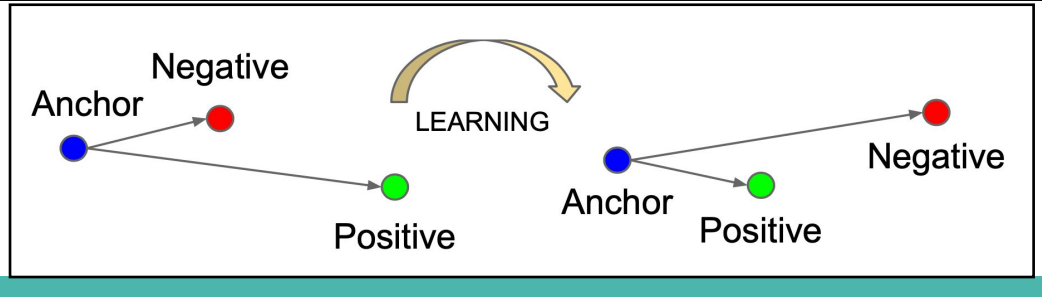
Contrastive Learning - Negative example - SimCLR

Distancia al ancla:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)}, \qquad (6)$$

- mínima para muestras positivas
- máxima para muestras negativas

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}$$



Contrastive Learning - Self-destillation

Similar al anterior, se prevén vistas de un “ancla”, pero sin vistas negativas. Suelen emplearse redes siamesas (dos redes de misma arquitectura con diferentes pesos). Dichas redes tienen un esquema diferente de actualización:

- online network: actualiza los pesos cada iteración, como de costumbre
- target network: se actualiza gradualmente mediante promediados del online network

Contrastive Learning - Self-destillation

"These models aim to maximize the similarity between two augmented versions of a single image while incorporating specific conditions to mitigate the risk of collapsing solutions."

La función objetivo a minimizar considera las salidas de las vistas generadas x1 , x2:

- ***p1 = g(f(x1))***
- ***z2 = f(x2)***

donde ***f( )*** es un encoder y ***g( )*** un "cabezal" MLP que procesa el espacio latente.

De esta forma se quiere minimizar la distancia coseno:

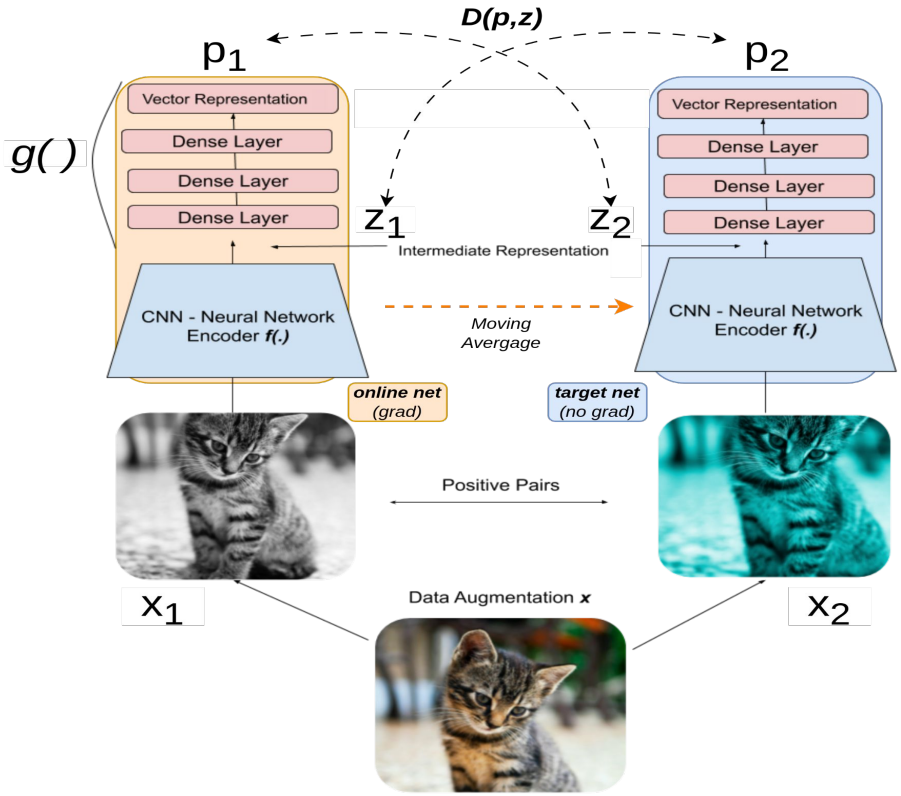
$$D\left(p_1, z_2\right) = -\frac{p_1}{\left\|p_1\right\|_2} \frac{z_2}{\left\|z_2\right\|_2}. \qquad (7)$$

Para que el entrenamiento sea simétrico, y que se considere el efecto online → target, se minimiza la función objetivo:

$$\mathcal{L} = \frac{1}{2}\left(D\left(p_1, stopgrad\left(z_2\right)\right) + D\left(p_2, stopgrad\left(z_1\right)\right)\right). \qquad (9)$$

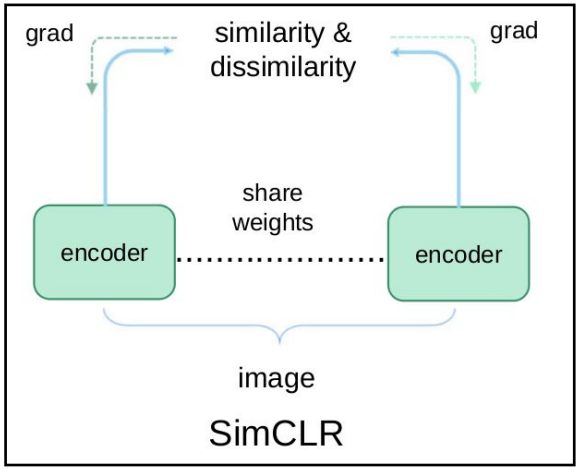
donde stopgrad( ) implica que la vista latente se trata como constante.

Contrastive Learning - Self-destillation-based CL

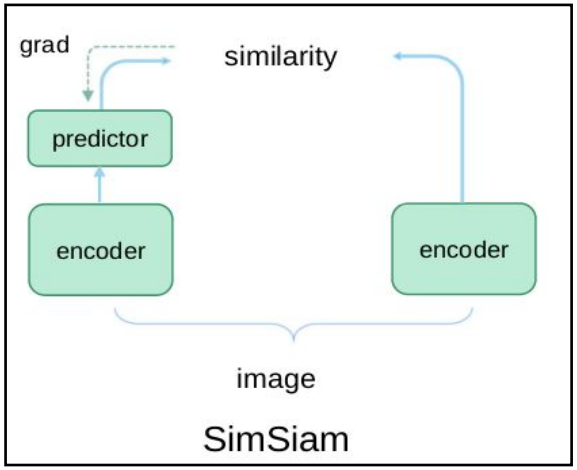


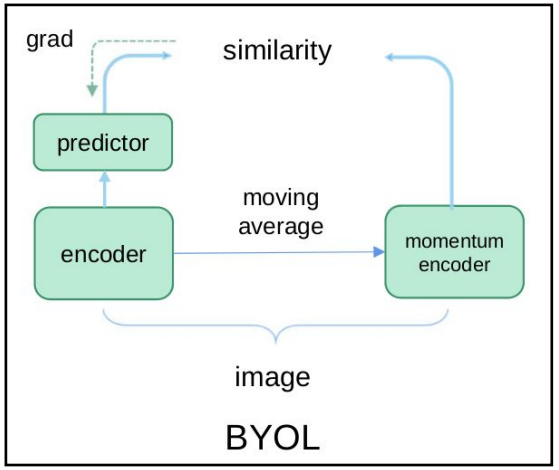
Contrastive Learning - Self-destillation-based

Negative Example



Self Distillation

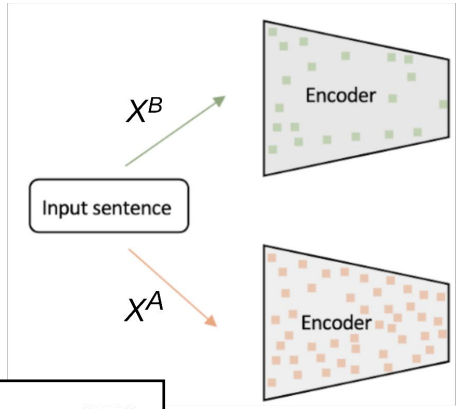


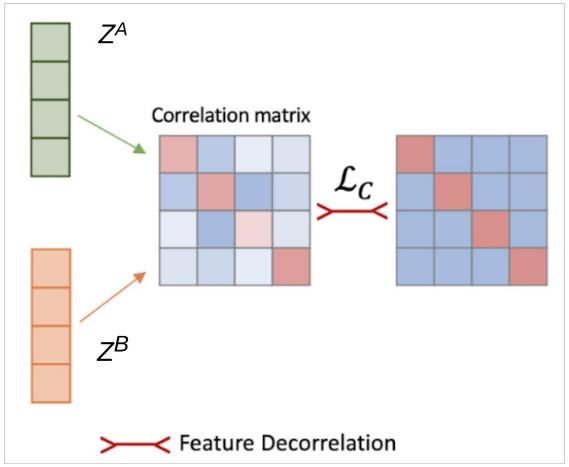


Contrastive Learning - Feature-Decorrelation

Barlow-Twins:

Operar sobre dos vistas (en batches) de las mismas instancias X^A y X^B





$$\mathcal{L}_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2. \qquad (10)$$

C_{ij} *matriz de correlación entre las salidas del encoder*

Contrastive Learning - Feature-Decorrelation - VICReg

Variance-Invariance-Covariance Regularization

*Se impone una **varianza de al menos gamma** en la dimensión del espacio latente. **S( )** desviación estándar + eps*

$$v\left(Z^A\right) = \frac{1}{d}\sum_{j=1}^{d}\max(0, \gamma - S\left(z_j^A, \varepsilon\right)). \qquad (12)$$

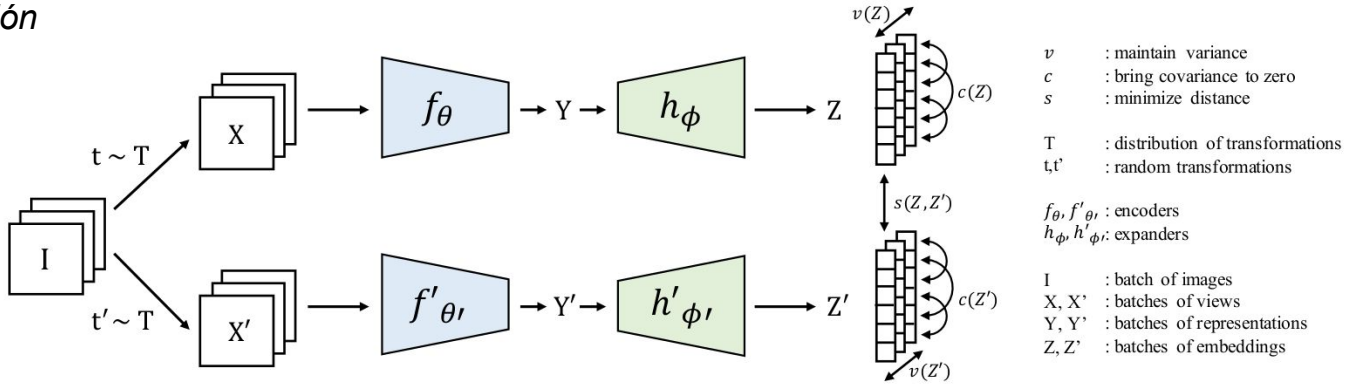


Figure 6: **VICReg**: penalizes variance, invariance, and co-variance terms to learn representations from unlabeled data.

Contrastive Learning - Feature-Decorrelation - VICReg

Variance-Invariance-Covariance Regularization

Asegurar cercanía entre embeddings *de la misma imagen para diferentes vistas*

$$s\left(Z^A, Z^B\right) = \frac{1}{n}\sum_{b=1}^{n}\left\|z_b^A - z_b^B\right\|_2^2. \qquad (14)$$

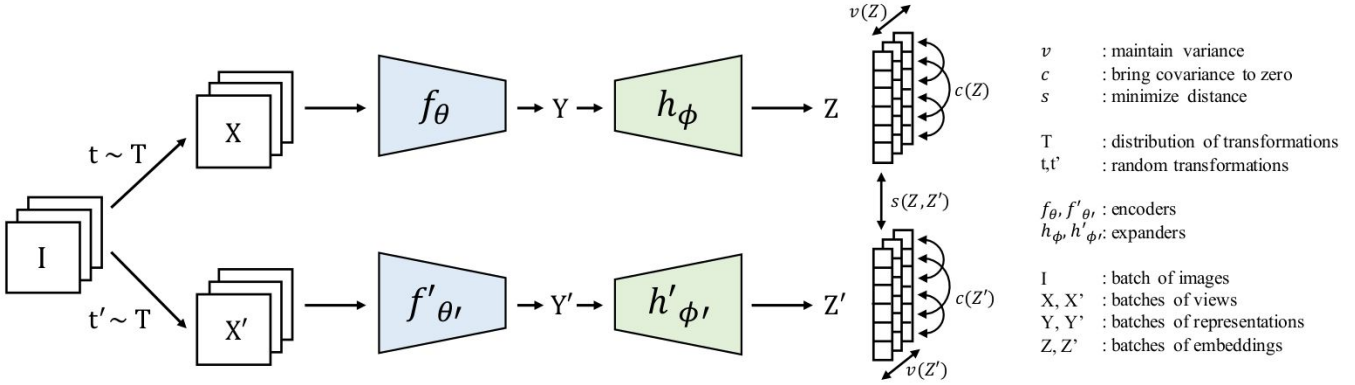


Figure 6: **VICReg**: penalizes variance, invariance, and co-variance terms to learn representations from unlabeled data.

Contrastive Learning - Feature-Decorrelation - VICReg

Variance-Invariance-Covariance Regularization

Correlación baja en embeddings para diferentes imagenes (ídem a Barlow Twins)

$$c(Z) = \frac{1}{d} \sum_{i \neq j} [C(Z)]_{i,j}^2 , \qquad (15)$$

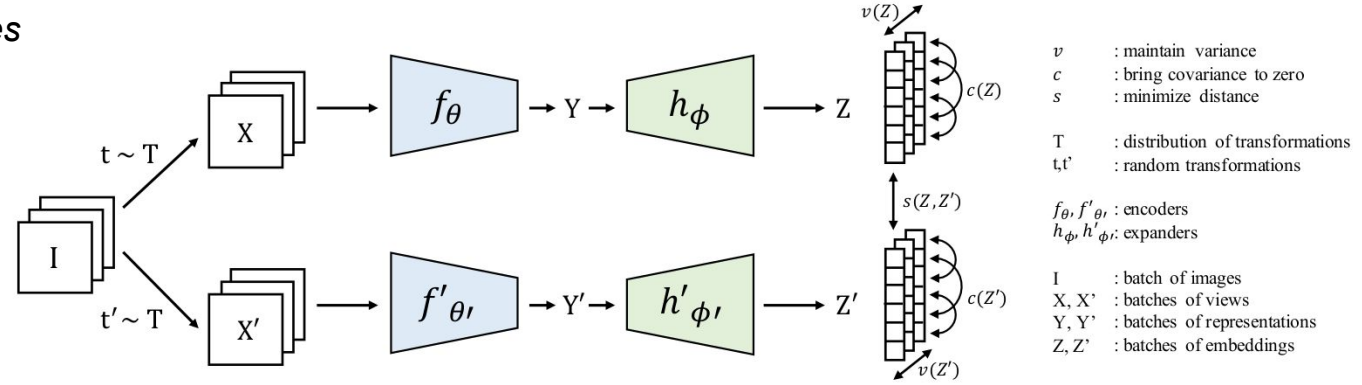


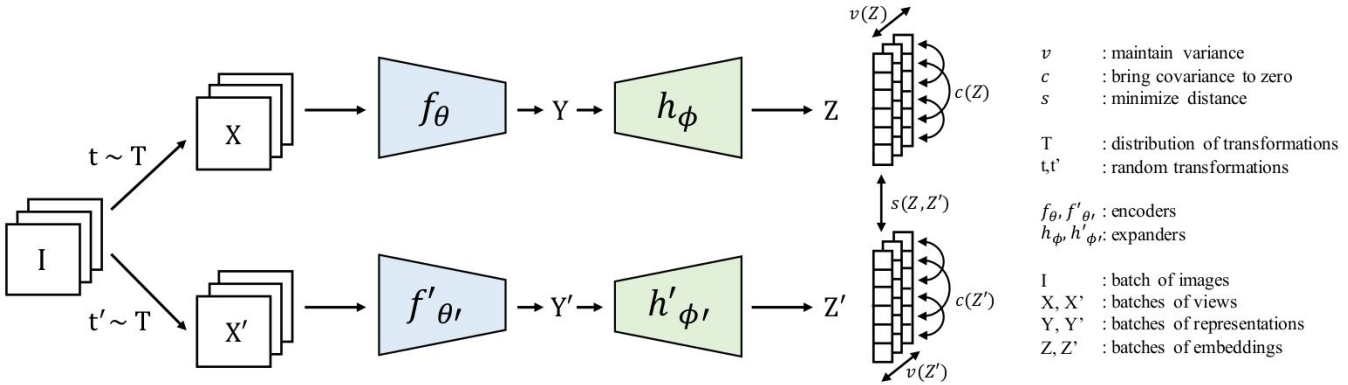
Figure 6: **VICReg**: penalizes variance, invariance, and co-variance terms to learn representations from unlabeled data.

Contrastive Learning - Feature-Decorrelation - VICReg

$$v\left(Z^A\right) = \frac{1}{d}\sum_{j=1}^{d}\max(0,\gamma - S\left(z_j^A,\varepsilon\right)).$$

$$s\left(Z^A,Z^B\right) = \frac{1}{n}\sum_{b=1}^{n}\left\|z_b^A - z_b^B\right\|_2^2.$$

$$c\left(Z\right) = \frac{1}{d}\sum_{i\neq j}\left[C(Z)\right]_{i,j}^2, \qquad (15)$$



v : maintain variance
c : bring covariance to zero
s : minimize distance

T : distribution of transformations
t,t' : random transformations

$f_\theta, f'_{\theta'}$: encoders
$h_\phi, h'_{\phi'}$: expanders

I : batch of images
X, X' : batches of views
Y, Y' : batches of representations
Z, Z' : batches of embeddings

$$\mathcal{L} = s\left(Z^A,Z^B\right) + \alpha\left(v\left(Z^A\right) + v\left(Z^B\right)\right) + \beta\left(C\left(Z^A\right) + C\left(Z^B\right)\right), \qquad (16)$$

Variance-Invariance-Covariance Regularization

- ***Varianza de al menos gamma*** *en la dimensión del espacio latente.* ***S( )*** *desviación estándar + eps*
- ***Asegurar cercanía entre embeddings*** *de la misma imagen para diferentes vistas*
- ***Correlación baja en embeddings para diferentes instancias*** *(ídem a Barlow Twins)*