fondazione olitec

Image-based Joint-Embedding Predictive Architecture

Architettura Predittiva di Inclusione Congiunta basata su Immagini

Massimiliano Nicolini

FONDAZIONE OLITEC www.olimaint.tech

Sommario

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

i-JEPA

Image-based Joint-Embedding Predictive Architecture

L'architettura predittiva di inclusione congiunta per immagini propone una tecnica auto supervisionata che cerca di apprendere caratteristiche dell'immagine altamente semantiche senza fare affidamento su potenziamenti realizzati a mano. L'idea principale di questo articolo è fornire a un blocco di contesto la rappresentazione di vari blocchi target.

Questo metodo ha ottenuto una precisione top 1 di +7,9 rispetto ai codificatori automatici mascherati sul set di dati ImageNet-1k.

Tipologie di architetture autogestite

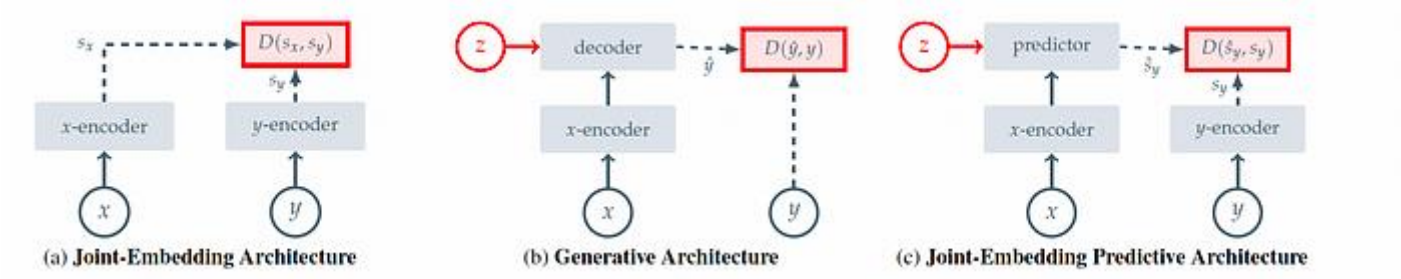


Figura 1: Diversi tipi di metodo autosupervisionato

In base al modo in cui le rappresentazioni delle immagini vengono confrontate e generate, i modelli autosupervisionati possono essere suddivisi in tre categorie:

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

1. **Architettura di incorporamento congiunto (JEA)**

Le architetture di incorporamento congiunto imparano a prevedere incorporamenti simili per input compatibili x, y e incorporamenti dissimili per input incompatibili. Uno di questi esempi è SimCLR, in cui diversi set di trasformazioni vengono applicati alla stessa immagine per generare coppie di input compatibili. Per il modello di ingressi non compatibili vengono utilizzate due immagini diverse. Nella Figura 1 l'encoder x e l'encoder y possono essere due encoder diversi oppure possono essere un unico encoder con meccanismo di condivisione del peso. La sfida principale della JEA è il collasso della rappresentazione, in cui il codificatore produce la stessa rappresentazione indipendentemente dall'input. Questo può essere risolto in alcuni modi, come codificatori di dimensioni diverse per il codificatore x e il codificatore y, separando esplicitamente gli incorporamenti di coppie negative.

2. **Architettura generativa**

Le architetture generative imparano a ricostruire un segnale y da un segnale compatibile x utilizzando una rete di decodificatori condizionata su variabili aggiuntive z. I codificatori automatici mascherati (MAE) ne sono un esempio. Nel caso di MAE, mascheriamo parte di un'immagine e passiamo solo le parti non mascherate dell'immagine attraverso il codificatore. Quindi nel decodificatore passiamo il token maschera (appreso durante l'addestramento e definisce quale parte viene ricostruita) per le parti mascherate insieme alle caratteristiche non mascherate apprese dal codificatore per prevedere la parte mascherata dell'immagine originale. Quindi l'immagine mascherata è l'input x, i token mascherati sono la condizione aggiuntiva z e la

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

parte mascherata dell'immagine originale è il segnale yche stiamo cercando di prevedere.

Il collasso della rappresentazione non è un problema qui.

3. **Architettura predittiva di inclusione congiunta (JEPA)**

Le architetture predittive di inclusione congiunta sono simili all'architettura generativa; l'unica differenza è che la perdita viene calcolata sullo spazio degli embedding e non sullo spazio degli input.
Le JEPA imparano a prevedere gli incorporamenti di un segnale y da un segnale compatibile x, utilizzando una rete di predittori condizionata su una variabile z aggiuntiva (possibilmente latente) per facilitare la previsione.
Un vantaggio di questo tipo di metodo rispetto alla JEA è che le JEPA non cercano rappresentazioni invarianti rispetto a una serie di aumenti di dati. In JEA, quando applichiamo due diversi set di trasformazioni, il modello è costretto ad apprendere che, dati i cambiamenti di rotazione, angolo e rumore, dovremmo ottenere gli stessi incorporamenti. JEPA tenta di prevedere più aree/regioni data una regione di contesto che dovrebbe forzare il modello ad apprendere più informazioni semantiche e non è necessario scegliere quali trasformazioni selezionare. Questo tipo di metodo soffre anche del collasso delle rappresentazioni. **Questo documento è una variante di questo tipo di metodo.** Risolvono il collasso della rappresentazione utilizzando encoder asimmetrici (encoder di diverse dimensioni).

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

Architettura predittiva di inclusione congiunta basata su immagini (I-JEPA)

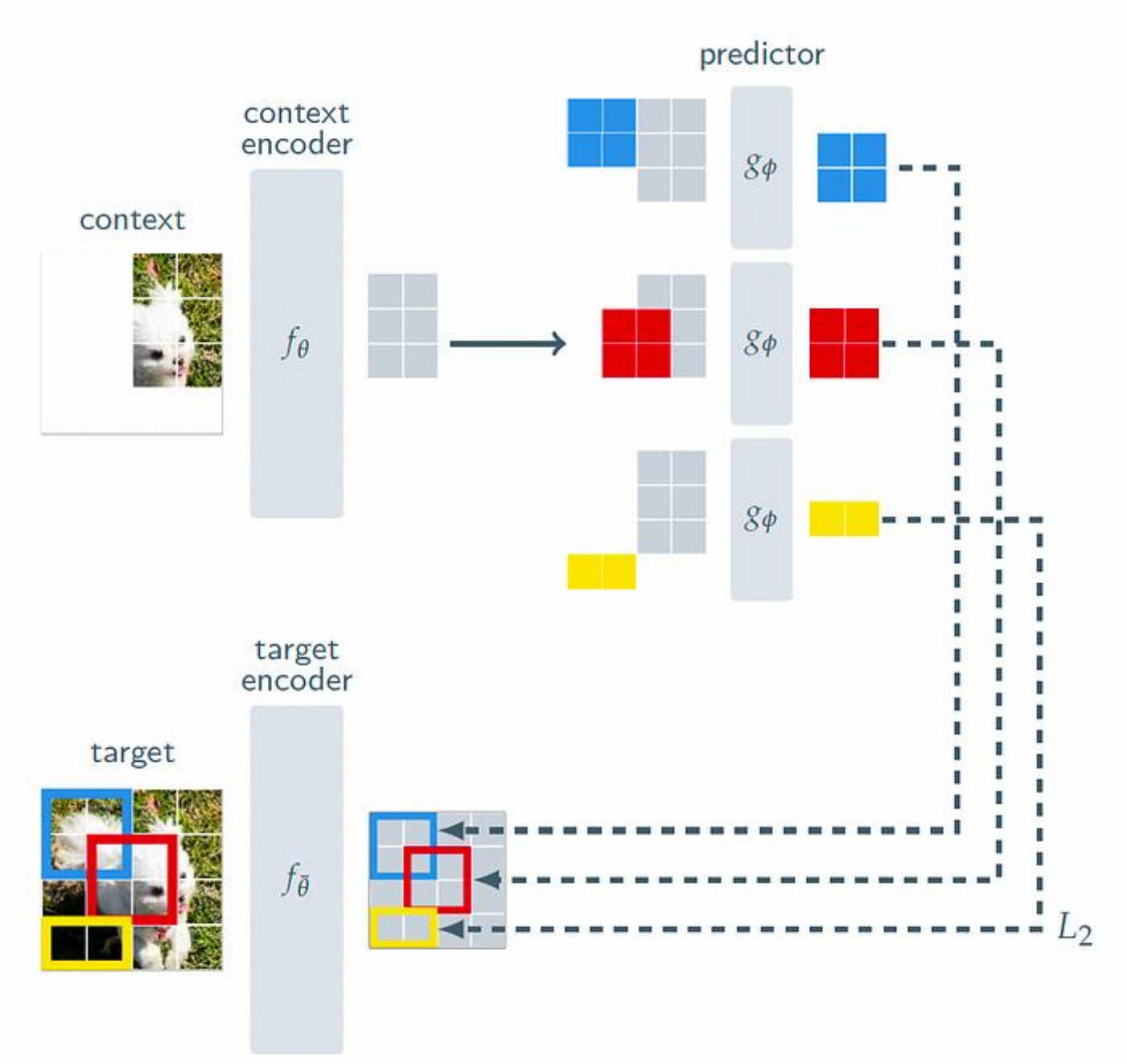


Figura 2: Architettura I-JEPA

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

Ci sono tre componenti principali nell'I-JEPA: selezione del contesto, selezione dell'obiettivo e predittore. Il contesto, i codificatori target e il predittore sono l'architettura Vision Transformer (ViT).

Selezione del contesto

Per contesto, gli autori selezionano un grande blocco quadrato casuale con una scala compresa tra (0,85, 1,0). Qui per scala si intende la dimensione rispetto all'immagine originale. Ciò significa che il blocco di contesto può coprire dall'85% dell'immagine all'immagine intera.

Questo blocco di contesto sarà composto anche da più patch. Le patch che si sovrappongono ai blocchi di destinazione vengono rimosse dal blocco di contesto **prima dell'elaborazione tramite il codificatore di contesto** . Il codificatore di contesto è un codificatore ViT.

Selezione del target

Gli obiettivi sono le rappresentazioni delle caratteristiche di diversi blocchi di immagini (o patch). Data un'immagine in input, questa viene convertita in una sequenza di patch non sovrapposte. Quindi vengono passati attraverso il codificatore di destinazione per ottenere le rappresentazioni delle caratteristiche. Ogni blocco dell'immagine avrà le rappresentazioni delle caratteristiche corrispondenti. Nella Figura 3 sono presenti 25 patch non sovrapposte sull'immagine originale. Quindi li passiamo attraverso il codificatore di destinazione per ottenere la rappresentazione delle caratteristiche dei 25 token.

Selezioniamo alcune delle rappresentazioni delle zone di quartiere per creare un obiettivo. Nella Figura 2, la rappresentazione del bersaglio rosso è composta da 4 rappresentazioni del patch e lo stesso per il bersaglio giallo. C'è una zona sovrapposta tra rosso e giallo.

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☏ +39 030 364332 int 5 +39 345 563 0496

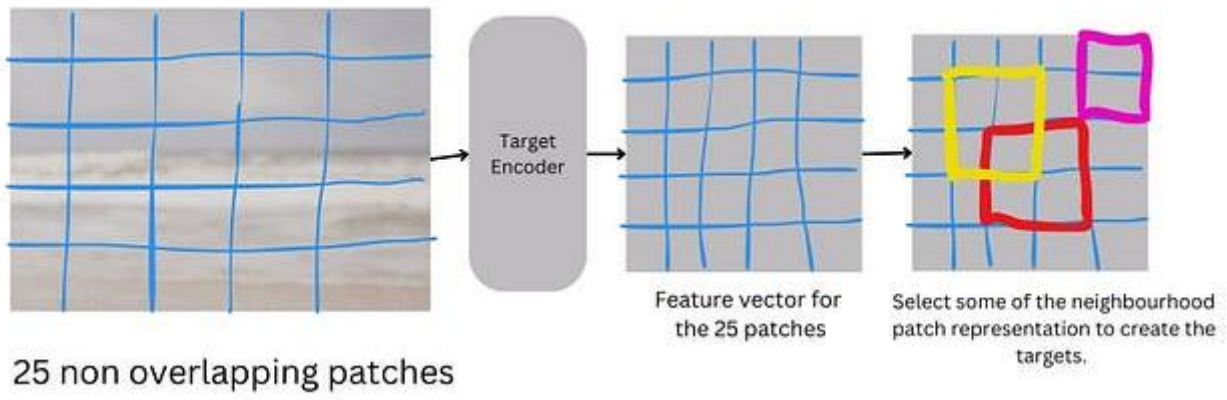


Figura 3: selezione del target dall'immagine di input

Gli autori hanno affermato che normalmente selezionano 4 blocchi target. Ogni blocco target è costituito da patch con proporzioni casuali $(0.75, 1.5)$e scala casuale $(0.15, 0.2)$. L'encoder di destinazione è un encoder ViT.

Esempi di contesto e strategia di mascheramento del target.

Predittore

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☏ +39 030 364332 int 5 +39 345 563 0496

Impariamo un token maschera durante l'addestramento (è simile ai token <SoS>e <EoS>nella PNL e viene appreso durante l'addestramento). Per ciascuna patch che vogliamo prevedere, aggiungiamo le codifiche posizionali corrispondenti a questo token maschera. Per la previsione del numero M di blocchi target, il predittore viene chiamato M volte con contesto codificato e (token maschera + incorporamenti posizionali) per tutte le patch come input.

L'architettura complessiva è mostrata nella Figura 2.

Perdita

La perdita è la distanza L2 tra le rappresentazioni a livello di patch previste e la rappresentazione a livello di patch di destinazione.

I pesi del codificatore di contesto vengono aggiornati attraverso i gradienti derivanti dalla perdita, ma i pesi del codificatore di destinazione vengono aggiornati tramite una media mobile esponenziale dei parametri del codificatore di contesto. In generale, il codificatore target della media mobile esponenziale funziona molto bene per l'addestramento dei JEA con Vision Transformers.

Risultati

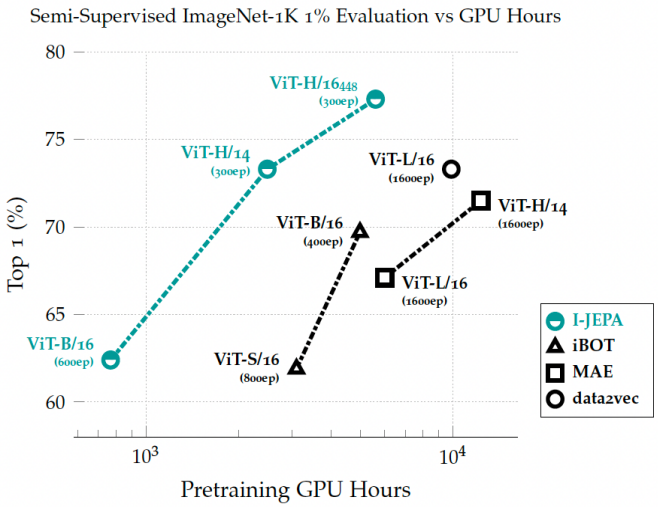
Per verificare che il modello abbia appreso buone caratteristiche semantiche dalle immagini durante il pre-addestramento, gli autori hanno riportato i risultati su vari compiti di classificazione delle immagini utilizzando i protocolli di sondaggio lineare e di messa a punto parziale.

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☏ +39 030 364332 int 5 +39 345 563 0496

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	77.3
MAE [35]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
CAE [21]	ViT-B/16	1600	70.4
	ViT-L/16	1600	78.1
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [20]	RN152 (2×)	800	79.1
DINO [17]	ViT-B/8	300	80.1
iBOT [75]	ViT-L/16	250	**81.0**



Risultati del sondaggio lineare.

Immagine a sinistra: i modelli vengono pre-addestrati sul set di dati ImageNet-1K, quindi viene addestrato un classificatore lineare sopra il codificatore congelato. Immagine a destra: le ore di pre-formazione per I-JEPA sono inferiori rispetto ad altri metodi.

Nel caso dell'addestramento semi-supervisionato (prima pre-addestramento, poi sondaggio lineare o messa a punto), questo documento produce risultati migliori. Per ciascun modello viene preso il meglio del sondaggio lineare e della messa a punto. Inoltre, il modello è ottimizzato sull'1% del set di dati ImageNet-1K.

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	73.3
MAE [35]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [75]	ViT-B/16	400	69.7
DINO [17]	ViT-B/8	300	70.0
SimCLR v2 [34]	RN151 (2×)	800	70.2
BYOL [34]	RN200 (2×)	800	71.2
MSN [3]	ViT-B/4	300	**75.7**

I-JEPA produce risultati migliori nell'attività di classificazione a valle con un numero inferiore di epoche (meno tempo di addestramento) sul set di dati ImageNet-1K

Valutazione lineare sui compiti di classificazione delle immagini a valle. I-JEPA supera significativamente le prestazioni dei metodi precedenti che non utilizzano miglioramenti (MAE e data2vec) e riduce il divario con i

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

migliori metodi basati sull'invarianza della vista che sfruttano i miglioramenti dei dati realizzati manualmente durante la pre-formazione.

Method	Arch.	CIFAR100	Places205	iNat18
Methods without view data augmentations				
data2vec [7]	ViT-L/16	81.6	54.6	28.1
MAE [35]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	**87.5**	**58.4**	**47.6**
Methods using extra view data augmentations				
DINO [17]	ViT-B/8	84.9	57.9	55.9
iBOT [75]	ViT-L/16	**88.3**	**60.4**	**57.3**

Trasferimento con sonda lineare per la classificazione delle immagini

Valutazione lineare su attività di basso livello a valle consistenti nel conteggio degli oggetti (Clevr/Count) e nella previsione della profondità (Clevr/Dist). Il metodo IJEPA cattura in modo efficace le caratteristiche dell'immagine di basso livello durante il preaddestramento e supera i metodi basati sull'invarianza della vista su attività quali il conteggio degli oggetti e la previsione della profondità.

Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [7]	ViT-L/16	85.3	71.3
MAE [35]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [17]	ViT-B/8	86.6	53.4
iBOT [75]	ViT-L/16	85.7	62.8

Trasferimento con sonda lineare per compiti di basso livello

Riteniamo inoltre che I-JEPA tragga vantaggio dalla formazione preliminare con set di dati più grandi.

Pretrain	Arch.	CIFAR100	Place205	INat18	Clevr/Count	Clevr/Dist
IN1k	ViT-H/14	87.5	58.4	47.6	86.7	72.4
IN22k	ViT-H/14	**89.5**	57.8	50.5	**88.6**	**75.0**
IN22k	ViT-G/16	**89.5**	**59.1**	**55.3**	86.7	73.0

Valutazione dell'impatto delle dimensioni del set di dati di pre-addestramento e delle dimensioni del modello sulle attività di trasferimento.

Perché funziona

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☎ +39 030 364332 int 5 +39 345 563 0496

Uno dei motivi per cui funziona è perché funziona sullo spazio latente rispetto allo spazio dei pixel. Quando si prevede nello spazio latente, ciascuno dei pezzi mancanti è una caratteristica o un oggetto di alto livello. Ciò ha molto più senso mascherare oggetti o caratteristiche di alto livello e provare a prevederli con un contesto perché questa idea di mascheramento proviene dalla PNL dove prevediamo un token/parola mascherato basato sui token/parole del vicinato. Nel dominio delle immagini, le parole dovrebbero essere mappate su oggetti, non su pixel, poiché i pixel non hanno molte informazioni semantiche individualmente come fanno le parole. Quando lo passiamo attraverso il codificatore, otteniamo funzionalità di alto livello dall'oggetto e le informazioni ridondanti a livello di pixel vengono generalmente rimosse. Questo è il motivo per cui funziona bene.

Olitec ®© Laboratorio di Ricerca e Sviluppo presso
Fondazione Olitec Caritate Christi - olimaint©® is a trade mark of Olimaint Company
Brescia Via Saleri 55/58 - Italy (EU)
Valmontone via Colle S.Angelo 2/O - Italy (EU)

www.olimaint.tech - desk@oliverso.it

☏ +39 030 364332 int 5 +39 345 563 0496