

**Borsa di studio attivata ai sensi di quanto disposto dal D.M. n. 1061 del 10/08/2021**

Titolo del progetto: EVER: vErso una formulazione VERde e sostenibile di algoritmi di auto-apprendimento di Rappresentazioni da video

La borsa sarà attivata sul seguente corso di dottorato accreditato per il XXXVII ciclo:  
INFORMATICA

Responsabile scientifico: Fabio Galasso

Area per la quale si presenta la richiesta: GREEN

Numero di mensilità da svolgere in azienda: 6

Numero di mensilità da svolgere all'estero: 6 presso EECS Department, University of California at Berkeley (UC Berkeley), Prof. Alexei Efros

Azienda: Panasonic R&D Company of America, California, USA

Il Dipartimento è disponibile a cofinanziare per un importo pari a euro: 10000

Dipartimento finanziatore: DIPARTIMENTO DI INFORMATICA con delibera del 21/09/2021

Progetto di ricerca:

L'auto-supervisione è un fattore chiave per il raggiungimento della prossima frontiera del machine learning scalabile e sostenibile, che ha l'obiettivo di macchine che esplorano autonomamente per accrescere la propria conoscenza, ad esempio dal web o da immagini e video acquisiti da sistemi mobili e robot. L'attuale "proiettile d'argento" per il progresso delle reti neurali profonde (Deep Neural Networks -- DNN) è stata la disponibilità di dati etichettati per addestrare i modelli DNN in modo supervisionato. Etichettare i dati significa annotare manualmente miliardi di fotogrammi dai video. Questo richiede molto tempo e ostacola la scalabilità dell'apprendimento a quantità di dati sempre crescenti, poiché l'etichettatura di un singolo frame può richiedere fino ad un'ora. Inoltre, questo è in conflitto con il principio stesso della possibile conoscenza di futuri sistemi e robot che dovrebbero essere in grado di apprendere (o continuare ad apprendere) da soli.

L'auto-supervisione si riferisce all'utilizzo della struttura stessa di immagini e video per supervisionare un algoritmo. In altre parole, l'algoritmo apprende prevedendo il pixel successivo in un'immagine o il fotogramma successivo nel video, dati alcuni pixel o fotogrammi iniziali. La supervisione è quindi "autonoma", perché fornita dal dato stesso, senza bisogno di etichettatura.

Puntare sull'auto-supervisione da video è un aspetto innovativo del progetto di dottorato proposto. La ricerca più recente ha considerato le immagini e l'ordine dei pixel all'interno delle immagini come auto-supervisione per l'apprendimento di modelli DNN. L'auto-supervisione dai video è ancora relativamente inesplorata e finora non si è ancora tradotta in progressi nell'apprendimento di rappresentazioni di video, ovvero in miglioramenti tangibili dell'accuratezza del modello addestrato dall'auto-supervisione. Ciò può essere spiegato dalla maggiore complessità dei video, rispetto all'immagine, per una serie di aspetti: i. c'è una mancata corrispondenza della frequenza di campionamento tra le dimensioni X e Y spaziali e la dimensione T temporale; ii. un punto nella posizione (x; y) al fotogramma t potrebbe non avere alcuna relazione con ciò che troviamo in quello stesso (x; y) al fotogramma t + k, poiché l'oggetto o la telecamera si saranno mossi in modo arbitrario (sebbene continuo) – problema detto anche "cosa è andato dove" [1].

Proponiamo di adottare un framework di apprendimento di tipo contrastive e di estendere l'apprendimento della somiglianza della coppia di immagini ai video mediante nuove tecniche per estrarre le corrispondenze temporali attraverso i fotogrammi video. L'apprendimento della rappresentazione auto-supervisionata da coppie corrispondenti di immagini, elaborate con data augmentation, è diventato efficace per l'apprendimento della rappresentazione basata su immagini [2]. Qui proponiamo di formulare il compito di trovare le corrispondenze temporali auto-supervisionate come un percorso su un grafo spazio-temporale. Il grafo è costruito da un video, in cui i nodi sono patch di immagini e solo i nodi nei fotogrammi vicini condividono un edge. La forza dell'edge è determinata dalla somiglianza secondo una rappresentazione appresa, il cui scopo è di stressare percorsi che collegano patch visivamente corrispondenti. L'apprendimento della rappresentazione si traduce nell'adattamento delle probabilità di transizione, per tenere traccia delle patch attraverso i fotogrammi nel grafo. Ispirati dai recenti progressi sull'auto-supervisione attraverso la coerenza temporale ciclica [3], forniamo l'auto-supervisione attraverso la formazione di video palindromi, come proposto anche da [4]. I video palindromi sono costruiti prendendo sequenze di fotogrammi in cui la prima metà viene ripetuta in ordine inverso. Ogni percorso spazio-temporale di patch a partire dal primo frame dovrebbe quindi tornare allo stesso punto di partenza.

Un aspetto verde innovativo di questa proposta di dottorato risiede nell'utilizzare oggetti invece che semplici patch quadrati estratti da immagini per l'apprendimento della rappresentazione auto-supervisionata nei video. I più recenti approcci di apprendimento auto-supervisionato hanno migliorato le prestazioni al costo di un enorme costo computazionale. In effetti, i metodi allo stato dell'arte richiedono potenza di calcolo di un ordine di grandezza in più rispetto all'addestramento supervisionato [5]. L'adozione di oggetti e di proprietà di questi per l'apprendimento di tipo contrastive è vantaggiosa perché ci sono molti meno oggetti rispetto ai pixel dell'immagine e alle patch nei fotogrammi video. Poiché il framework di path-finding basato su grafi proposto scala in maniera almeno quadratica con il numero di nodi, così facendo si riduce il costo di calcolo di almeno 5 cinque volte [6].

Ricerche recenti hanno illustrato l'impatto dell'efficienza computazionale degli algoritmi DNN sull'ambiente in termini di emissione di CO2 equivalente. Un numero crescente di articoli sta ponendo l'attenzione sull'impatto della memoria richiesta e dei tempi di addestramento dei modelli [7,8], evidenziando che alcune moderne tecniche DNN emettono una quantità di CO2 cinque volte superiore alle emissioni di un'auto americana media durante tutta la sua vita (compresi carburante e la costruzione dell'auto stessa) [9].

Date queste premesse, una seconda innovazione verde della proposta di dottorato è la riformulazione del framework di apprendimento della rappresentazione basata su video in termini delle recenti reti neurali convoluzionali a grafo (Graph Convolutional Networks -- GCN). Tale riformulazione viene abilitata dall'uso di oggetti, entità semantiche la cui interazione è modellabile tramite grafo. La ricerca più recente del nostro gruppo ha raggiunto una svolta nelle prestazioni sul task di previsione della posa umana mediante un nuovo GCN separabile nello spazio e tempo, detto STS-GCN [10]. Se confrontato con le migliori performance di altre tecniche sui tre più grandi e recenti benchmark, il nostro modello riduce l'errore di previsione della posa di almeno il 33%, adottando solo una frazione dei parametri del modello, il 2,5%, e di conseguenza si corrisponde tempo di addestramento. Ipotizziamo che una riformulazione basata su GCN dell'apprendimento della rappresentazione video auto-supervisionato migliorerebbe le prestazioni e ridurrebbe l'impronta di carbonio dell'addestramento del modello DNN, come ci proponiamo di ricercare e dimostrare nel corso di questo progetto di dottorato.

[1] Josh Wills, Sameer Agarwal, and Serge Belongie (2003). "What went where". In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In Proc. International Conference on Machine Learning (ICML'20)

[3] Xiaolong Wang, Allan Jabri, and Alexei A Efros (2019). "Learning correspondence from the cycle-consistency of time". In Proc. Computer Vision and Pattern Recognition (CVPR'19).

[4] Allan Jabri, Andrew Owens, Alexei A. Efros (2020). "Space-Time Correspondence as a Contrastive Random Walk". In Proc. Conference on Neural Information Processing Systems (NeurIPS'20)

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. "Big self-supervised models are strong semi-supervised learners". In Proc. Conference on Neural Information Processing Systems (NeurIPS'20)

[6] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, João Carreira (2021). "Efficient Visual Pretraining with Contrastive Detection". Pre-print ArXiv'21

[7] Emma Strubell, Ananya Ganesh, Andrew McCallum (2020). "Energy and Policy Considerations for Modern Deep Learning Research". In Proc. AAAI Conference on Artificial Intelligence (AAAI'20)

[8] Titouan Parcollet, Mirco Ravanelli (2019). "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers". Interspeech 2021

[9] Emma Strubell, Ananya Ganesh, Andrew McCallum (2019). "Energy and policy considerations for deep learning in NLP." In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL'19)

[10] Theodoros Sofianos, Alessio Sampieri, Luca Franco and Fabio Galasso (2021). "Space-Time-Separable Graph Convolutional Network for Pose Forecasting". In Proc. International Conference on Computer Vision (ICCV'21)

Titolo del progetto (inglese): EVER: towards grEen and sustainable large-scale self-supervised Video rEpresentation LeaRning

Progetto di ricerca (inglese):

Self-supervision stands as the key enabler to the next frontier of scalable machine learning which can sustainably grow: the one of machines which scout for their knowledge autonomously, from the web or from images and videos acquired from mobiles and robots. The current silver bullet for the progress of Deep Neural Networks (DNN) has been the availability of labelled data to train the DNN models supervisedly. Labelling data means manually annotating billions of frames from videos. This is time-consuming and it hinders the scalability of machine learning to ever growing amounts of data, as labelling a single frame may take up to one hour. Furthermore, this conflicts with the very principle of future systems and robots which should be capable to learn (or continue learning) by themselves.

Self-supervision refers to using the very structure of images and videos to supervise an algorithm. In other words, the algorithm learns by predicting the next pixel in an image or the next frame in the video, given a few initial pixels or frames. The supervision is therefore "self", because it is provided by the very piece of data, without any need for labelling.

Focusing on self-supervision from videos is an innovative aspect of the proposed PhD project. Most recent research has considered images and the pixel order within images as the self-supervision for the DNN model. Self-supervision from videos is still relatively unexplored and it has so far not yet translated into advances in machine representation learning, i.e. into tangible improvements of accuracy from the self-supervisedly trained model. This may be explained by the added complexity of videos, compared to image, for a number of aspects: i. there is a sampling rate mismatch between the spatial X and Y Vs. the temporal T dimensions; ii. a physical point depicted at position (x; y) in frame t

might not have any relation to what we find at that same  $(x; y)$  in frame  $t + k$ , as the object or the camera will have moved in arbitrary (albeit smooth) ways -- also termed the "what went where" issue [1].

We propose to adopt a contrastive learning framework and to extend the image-pair similarity learning to videos by novel techniques for mining temporal correspondences across the video frames. Self-supervised representation learning from matching pairs of data-augmented images have become effective for image-based representation learning [2]. Here we propose to formulate the task of self-supervisedly finding temporal correspondences as a pathfinding on a space-time graph. The graph is constructed from a video, where nodes are image patches and only nodes in neighboring frames share an edge. The strength of the edge is determined by similarity under a learned representation, whose aim is to place weight along paths linking visually corresponding patches. Learning the representation translates to fitting transition probabilities, to track the patches across the frames in the graph. Inspired by recent progress on self-supervision via cycle-consistency of time [3], we provide self-supervision by training on palindrome videos, as also proposed by [4]. Palindrome videos are constructed by taking sequences of frames where the first half is repeated backwards. Every Spatio-temporal path of patches starting on the first frame should therefore return to its starting point.

An innovative green aspect of this PhD proposal lies in the proposition of using objects instead of plain square image patches for self-supervised representation learning in videos. Most recent self-supervised learning approaches have improved performance at the cost of a tremendous computational cost. In fact, state-of-the-art methods require an order of magnitude more computation than supervised pretraining [5]. Adopting objects and object-level features for contrastive learning is advantageous because there are much fewer objects than image pixels and patches in the video frames. Since the proposed graph-based pathfinding framework scales at least quadratically with the number of nodes, doing so reduces the computation cost of at least 5 five times [6].

Recent research has illustrated the impact of computational efficiency of DNN algorithms on the environment in terms of equivalent CO<sub>2</sub> emission. An increasing number of articles are posing the attention on the impact of the required memory and training times of the models [7,8], with some techniques emitting five times as much CO<sub>2</sub> as the lifetime emissions of the average American car (including fuel and the car manufacturing itself [9]).

Given these premises, a second green innovation of this PhD proposal is the reformulation of the video-based representation learning framework in terms of the recent graph convolutional neural networks (GCN). This reformulation is enabled by the proposed use of objects, i.e. both the objects as semantic entities and their interaction, which may be modelled by a graph. Most recent research from our group has achieved a breakthrough in performance in the task of human pose forecasting by a novel space-time separable GCN, STS-GCN [10]. When benchmarked on three large-scale pose forecasting datasets, our model reduces the pose forecasting error by at least 33%, while only adopting a fraction of the model parameters, 2.5%, and the corresponding the training time. We speculate that a GCN-based reformulation of self-supervised video representation learning would both improve performance and reduce the carbon footprint of training the DNN model, as we plan to research during the proposed PhD project.

[1] Josh Wills, Sameer Agarwal, and Serge Belongie (2003). "What went where". In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In Proc. International Conference on Machine Learning (ICML'20)

[3] Xiaolong Wang, Allan Jabri, and Alexei A Efros (2019). "Learning correspondence from the cycle-consistency of time". In Proc. Computer Vision and Pattern Recognition (CVPR'19).

- [4] Allan Jabri, Andrew Owens, Alexei A. Efros (2020). "Space-Time Correspondence as a Contrastive Random Walk". In Proc. Conference on Neural Information Processing Systems (NeurIPS'20)
- [5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. "Big self-supervised models are strong semi-supervised learners". In Proc. Conference on Neural Information Processing Systems (NeurIPS'20)
- [6] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, João Carreira (2021). "Efficient Visual Pretraining with Contrastive Detection". Pre-print ArXiv'21
- [7] Emma Strubell, Ananya Ganesh, Andrew McCallum (2020). "Energy and Policy Considerations for Modern Deep Learning Research". In Proc. AAAI Conference on Artificial Intelligence (AAAI'20)
- [8] Titouan Parcollet, Mirco Ravanelli (2019). "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers". Interspeech 2021
- [9] Emma Strubell, Ananya Ganesh, Andrew McCallum (2019). "Energy and policy considerations for deep learning in NLP." In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL'19)
- [10] Theodoros Sofianos, Alessio Sampieri, Luca Franco and Fabio Galasso (2021). "Space-Time-Separable Graph Convolutional Network for Pose Forecasting". In Proc. International Conference on Computer Vision (ICCV'21)