

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA
Dipartimento di Ingegneria “Enzo Ferrari”
Corso di Laurea Magistrale in Ingegneria Informatica (D.M. 270/04)

Relatore:

Prof.ssa Sonia Bergamaschi

Correlatore:

Pietro Leo

Candidato:

Matteo Gabrielli

**ACTION RECOGNITION PER STIMARE
LE ACTIVITIES OF DAILY LIVING (ADL)
DI PERSONE ANZIANE**

Anno Accademico

2017 - 2018



OUTLINE

1. Analisi use-case fornito da IBM Italia e «IRCCS Casa Sollievo della Sofferenza»
2. Scelta del dataset, estrazione e pre-processing dei dati
3. Creazione del modello usando un approccio di Auto ML
4. Creazione del modello tramite Transfer Learning
5. Presentazione dei risultati e sviluppi futuri



MULTIDIMENSIONAL PROGNOSTIC INDEX

«*Il Multidimensional Prognostic Index (MPI) è un indice prognostico di mortalità ad un anno basato su informazioni ottenute da una Valutazione Multidimensionale (VMD) del soggetto anziano*»

MPI

- 8 domini, 63 elementi
- La conoscenza prognostica dell'outcome del paziente anziano può indirizzare o influenzare le scelte diagnostiche e terapeutiche.

Activities of Daily Living

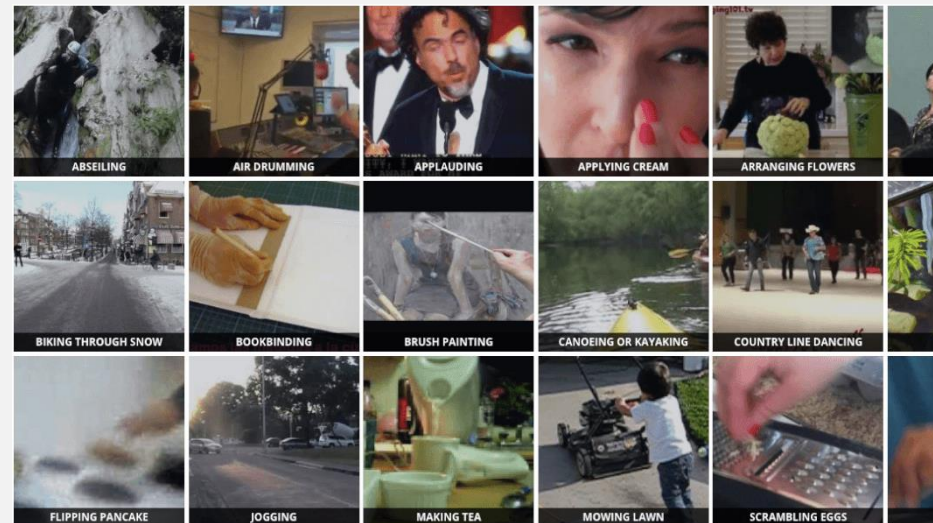
Le ADL sono attività di base eseguite durante la vita di un essere umano: lavarsi, mangiare, vestirsi, ...



OBIETTIVI DEL PROGETTO

Progettare e sviluppare un modello di Deep Learning che sia in grado di riconoscere azioni svolte in un video contenente scene di vita quotidiana, svolte da un soggetto anziano all'interno dell'abitazione.

I risultati del modello sono a supporto della compilazione dell'ADL e IADL.



SCELTA DEI DATI: «MOMENTS IN TIME»

La generazione dei dati è molto onerosa in termini di tempo e costi

È stata eseguita un'analisi dei vari dataset disponibili per la Action Recognition per la ricerca di una fonte di dati pronta e validata.

Il dataset «Moments in Time» ha:

- 339 classi
- 1 milione di video di 3 secondi con label
- ampia variazione intra-classe dei video

ABOUT TEAM DEMO PAPER CHALLENGE DOWNLOAD EXPLORE

Moments in Time Dataset

A large-scale dataset for recognizing and understanding action in videos

Moments is a research project in development by the MIT-IBM Watson AI Lab. The project is dedicated to building a very large-scale dataset to help AI systems recognize and understand actions and events in videos.

Today, the dataset includes a collection of one million labeled 3 second videos, involving people, animals, objects or natural phenomena, that capture the gist of a dynamic scene.

- MOMENTS**
Three seconds events capture an ecosystem of changes in the world: 3 seconds convey meaningful information to understand how agents (human, animal, artificial or natural) transform from one state to another.
- DIVERSITY**
Designed to have large inter-class and intra-class variation that represent dynamical events at different levels of abstraction (i.e. "opening" doors, drawers, curtains, presents, eyes, mouths, and even flower petals).
- GENERALIZATION**
A large-scale, human-annotated video dataset capturing visual and/or audible actions, produced by humans, animals, objects or nature that together allow for the creation of compound activities occurring at longer time scales.
- TRANSFERABILITY**
Supervised tasks on a large coverage of the visual and auditory ecosystem help construct powerful but flexible feature detectors, allowing models to quickly transfer learned representations to novel domains.

DATA EXTRACTION & CLEANING - PIPELINE

Linguaggio utilizzato:



Interazione con i video:



	A	B	C	D	E	F	G
1	CATEGORIES					CATEGORIES MERGED	CATEGORIES (final name)
2	calling					calling/telephoning	telephoning
3	cooking					cooking	cooking
4	dining					dining/eating	eating
5	dressing					dressing	dressing
6	drinking					drinking	drinking
7	dusting					dusting/mopping/vacuuming	housekeeping
8	eating					reading/studying/writing	reading-writing
9	mopping					sewing	sewing
10	reading					smoking	smoking
11	sewing					socializing	socializing
12	smoking					standing	standing
13	socializing						
14	standing						
15	studying						
16	telephoning						
17	vacuuming						
18	writing						

Scelta delle classi (11)

Estrazione dei dati

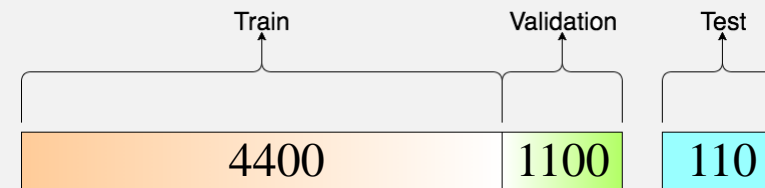
Pre-processing dei video

Estrazione frames dai video

Data cleaning manuale sulle immagini

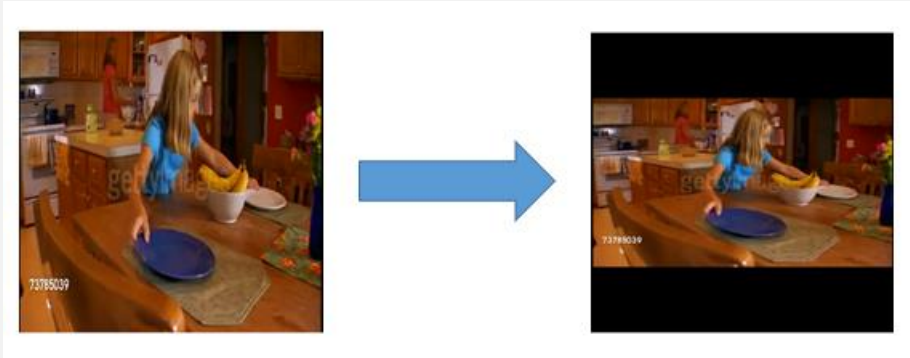
Creazione del dataset,
500 immagini per classe

500 x 11 = 5500 immagini totali

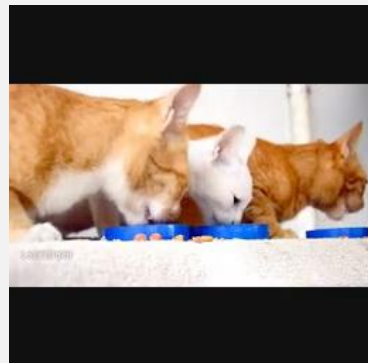


DATA PRE-PROCESSING

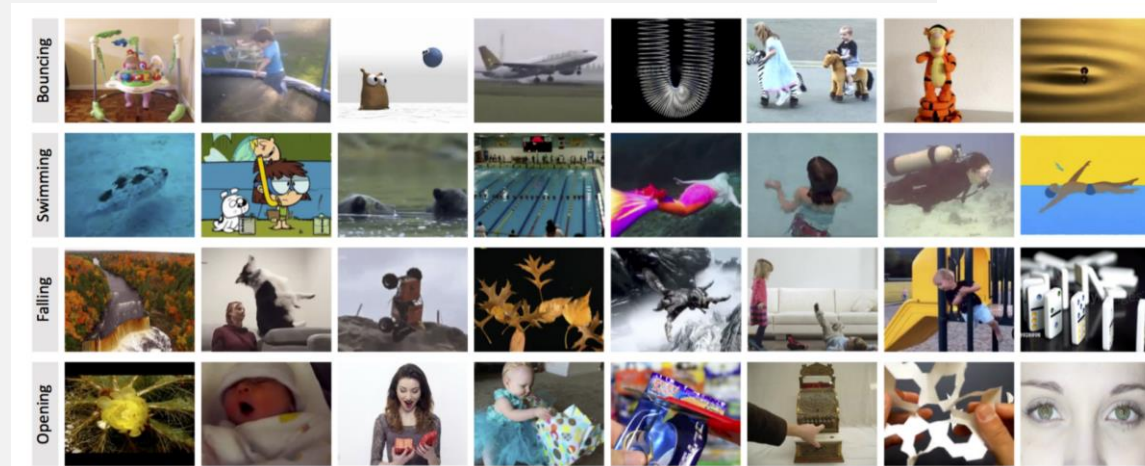
«La scarsa qualità dei dati è il nemico numero uno per l'uso diffuso e redditizio dell'apprendimento automatico.»



Garantire la correttezza dell'immagine



Immagini che non rispecchiano il nostro caso



Interpretazione del nome dell'azione

The background is an abstract, textured composition of various colors including red, yellow, green, blue, and purple, resembling a marbled or painted surface. A dark, semi-transparent overlay covers the entire image, providing a high-contrast background for the text.

MODELLI DI DEEP LEARNING

*Che performance è
possibile raggiungere?*

APPROCCIO AUTO ML - NEUNETS

Sintesi automatica di modelli di Reti Neurali

1. Forniti in input i dati
2. NeuNetS crea automaticamente la rete neurale più adatta per i dati forniti ed effettua il training
3. Risultati forniti in ~2 ore

Punti di forza:

- Accessibile gratuitamente in IBM Cloud
- Nessuna riga di codice per avere un primo risultato

The screenshot displays the IBM Watson Studio interface for a project named 'moments-11-all'. The top navigation bar includes the IBM logo, an 'Upgrade' button, and a user profile icon. Below the navigation bar, the project name and a breadcrumb trail are visible. A progress bar shows the job status: 'Job accepted', 'Preprocessing', 'Synthesizing', and 'Completed'. The 'Synthesizing' step is currently active. To the right of the progress bar are buttons for 'Download model' and 'Deploy model to Watson Machine Learning'. The main content area is divided into two panels. The left panel, titled 'moments-11-all', shows the 'Status' as 'Synthesizing complete. You can download or deploy your model.' Below this, the 'Performance' metrics are listed: Accuracy 45.5%, Precision 0.469, and Recall 0.456. The 'Training data' section includes a table with the following data:

Content type	image
Source bucket	training-11-all
Number of classes	11

The right panel, titled 'Label statistics' and 'Confusion matrix', displays a table of the 5 most correct labels. The table has columns for 'Actual' (standing, readin..., cooking, housek..., teleph...) and 'Actual totals'. The rows represent the predicted labels. The diagonal elements (top-left to bottom-right) are shaded, indicating correct classifications. The 'Actual totals' column shows the number of instances for each label. The table data is as follows:

Predicted \ Actual	standing	readin...	cooking	housek...	teleph...	Actual totals
standing	0.60	0.04	0.02	0.06	0.08	1006
readin...	0.11	0.43	0.03	0.04	0.13	1236
cooking	0.08	0.05	0.59	0.05	0.03	868
housek...	0.18	0.05	0.04	0.55	0.05	893
teleph...	0.10	0.14	0.02	0.03	0.51	964
Predicted totals	1662	1102	775	846	1165	8575

At the bottom right of the confusion matrix panel, it says 'Showing 5 of 11 total classes' and provides links for 'Show full confusion matrix' and 'Download all'.

MODELLO ALLENATO SU «MOMENTS IN TIME»

Modello di partenza per effettuare il Transfer Learning



Top-5 Actions:

31.5% -> dining

4.0% -> feeding

3.9% -> eating

2.8% -> serving

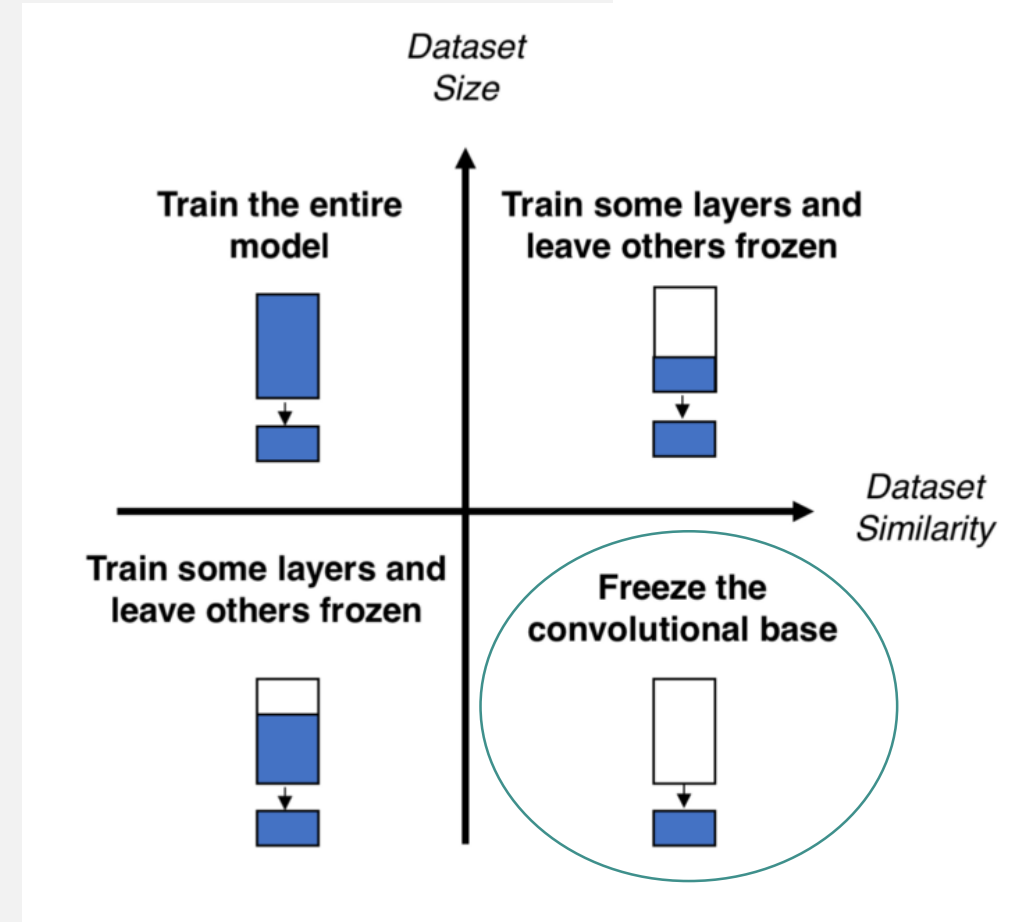
2.6% -> socializing



Class Activation Mapping

Model	Modality	Top-1 (%)	Top-5 (%)
Chance	-	0.29	1.47
ResNet50-scratch	Spatial	23.65	46.73
ResNet50-Places	Spatial	26.44	50.56
ResNet50-ImageNet	Spatial	27.16	51.68
TSN-Spatial	Spatial	24.11	49.10
BNInception-Flow	Temporal	11.60	27.40
TSN-Flow	Temporal	15.71	34.65
SoundNet	Auditory	7.60	18.00
TSN-2stream	Spatial+Temporal	25.32	50.10
TRN-Multiscale	Spatial+Temporal	28.27	53.87
I3D	Spatial+Temporal	29.51	56.06
Ensemble (SVM)	S+T+A	31.16	57.67

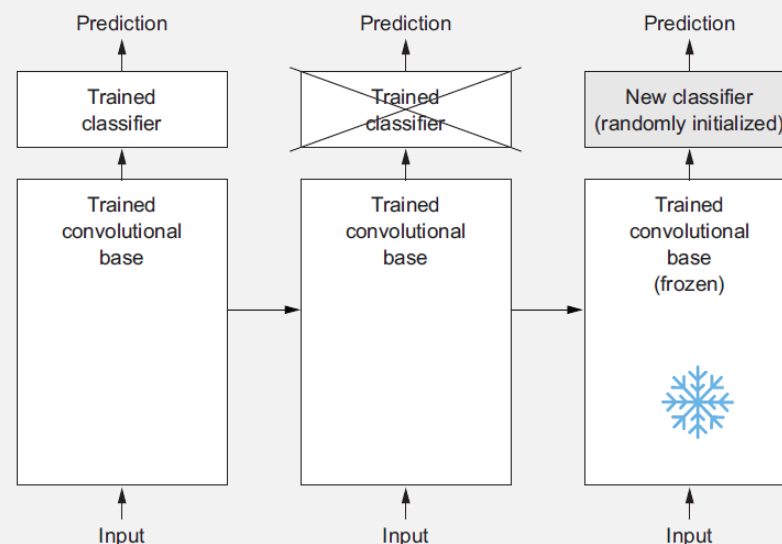
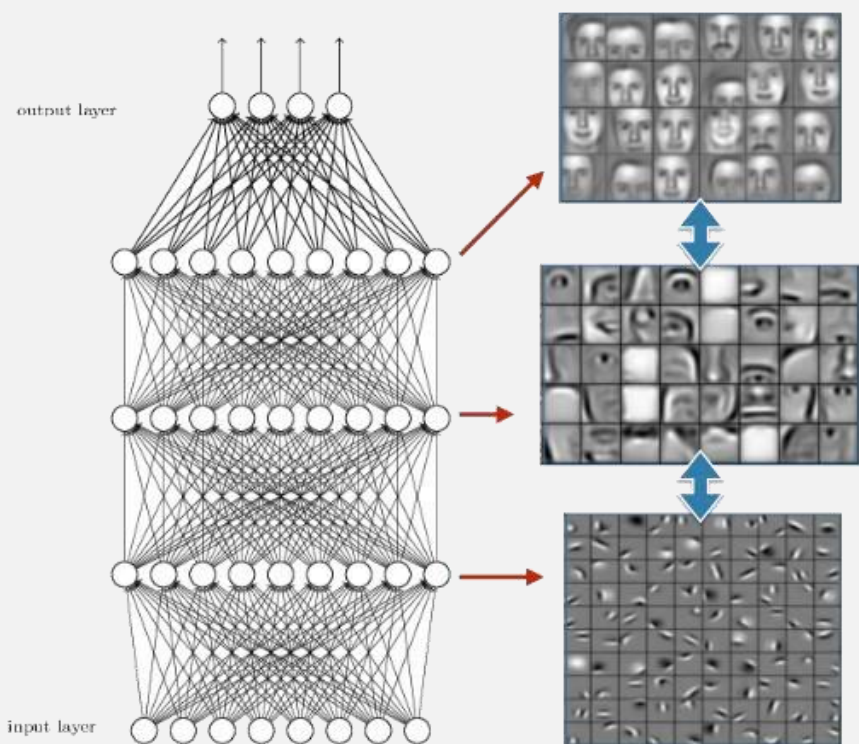
TABLE 1: Classification Accuracy: We show Top-1 and Top-5 accuracy of the baseline models on the validation set.



TRANSFER LEARNING

“Transfer learning is a popular method in computer vision because it allows us to build accurate models in a timesaving way” (Rawat & Wang, 2017).

“The applications of skills, knowledge and/or attitudes that were learned in one situation to another learning situation (Perkins, 1992)”



Training su Kaggle Kernel:

- GPU NVIDIA Tesla K80
- ~ 1 minuto ogni epoca (un iterazione su tutto il dataset)




PYTORCH



kaggle

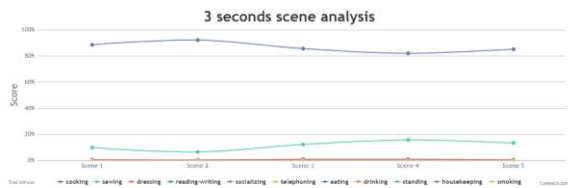
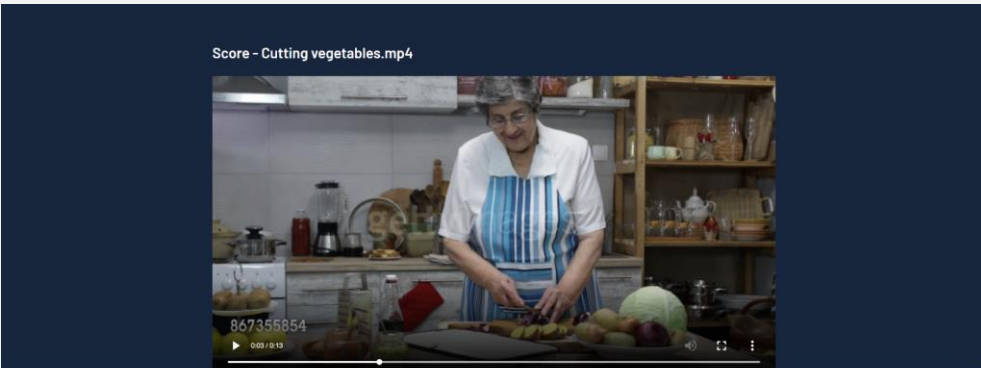
RISULTATI SUL DATASET DI TEST (110 ELEMENTI)

	NeuNetS su 500 immagini per classe	NeuNetS su tutte le immagini di una classe	Transfer learning sul modello Moments 	Transfer learning sul modello ImageNet
Accuracy	18.18 %	36.36%	64.55%	48.18%
Cooking	70.0 %	70.0 %	90.0 %	90.0 %
Dressing	0.0 %	0.0 %	60.0 %	50.0 %
Drinking	20.0 %	70.0 %	100.0 %	70.0 %
Eating	10.0 %	50.0 %	80.0 %	50.0 %
Housekeeping	50.0 %	80.0 %	90.0 %	100.0 %
Reading-writing	10.0 %	40.0 %	60.0 %	50.0 %
Sewing	0.0 %	0.0 %	50.0 %	20.0 %
Smoking	30.0 %	20.0 %	30.0 %	10.0 %
Socializing	0.0 %	30.0 %	50.0 %	20.0 %
Standing	0.0 %	0.0 %	0.0 %	0.0 %
Telephoning	10.0 %	40.0 %	100.0 %	70.0 %

WEB-APP: DEMO FOR MODEL INTERACTION



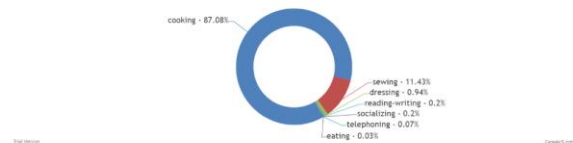
IBM Cloud



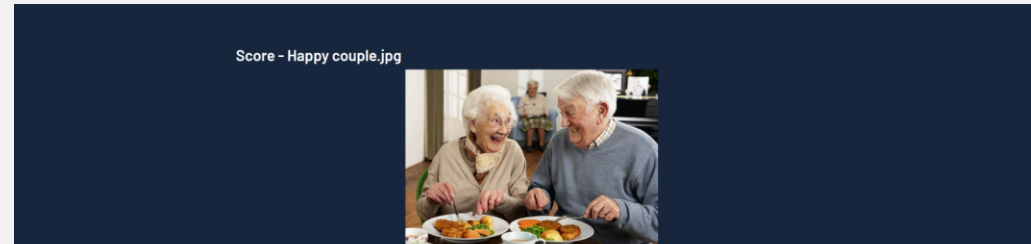
Transfer Learning from Moments pretrained

Action predicted: cooking at 87.1%

Predictions



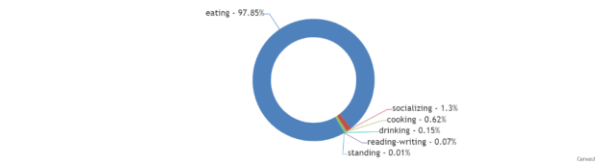
*Test modello su video
(evoluzione temporale delle scene)*



Transfer Learning from Moments pretrained

Action predicted: eating at 97.8%

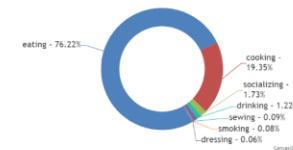
Predictions



Transfer Learning from ImageNet pretrained

Action predicted: eating at 76.2%

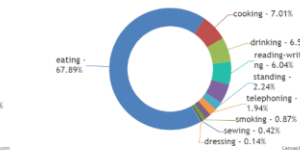
Predictions



NeuNetS model

Action predicted: eating at 67.9%

Predictions



Test modelli su immagine

SVILUPPI FUTURI

- Da Multi-Class a Multi-Label (concorrenza azioni)
- Integrare la componente audio e quella temporale al modello (solo componente spaziale)
- Migliorare l'accuracy del modello collezionando più dati e lavorando sui parametri che influiscono sul training del modello
- Integrare soluzioni di Object Detection per identificare gli oggetti nella scena
- Estendere il numero di azioni riconoscibili
- Cominciare una fase sperimentale per poter raccogliere feedback e dati sul campo

GRAZIE

Matteo Gabrielli  <https://github.com/mattegab13>