**Discorso Presentazione Deep Learning**

**PARTE DI GIO [x minuti totali] Tempi: 9:26, 9:05, 8:44, 9:01, 8:57,**

**Slide 1: Titolo Slide [10 sec]**

**Buongiorno. Il nostro progetto prende il nome di Video Classification: Human Action Recognition on HMDB51 database.**

**Slide 2: Task: Human action recognition in videos [1:40 minuti slide, 1:50 minuti totale]**

**HMDB51** è un grande database di video per l’human action recognition, ovvero il task principale del nostro progetto.  ~~Con Human Action Recognition (~~**~~HAR~~**~~) si identifica quel task~~ che prevede la classificazione di una o più azioni eseguita da uno o più soggetti umani. A **differenza** di altri task di classificazione, questa necessita una serie di data points campionati nel tempo al fine di poter predire correttamente l’azione eseguita ~~e per questo il task può essere visto come un problema di classificazione di serie temporali.~~ Un **approccio tradizionale** al problema è quello di utilizzare sensori indossabili ~~(come accelerometri e giroscopi)~~ al fine di generare i data points~~, il metodo è efficace,~~ però, l’impossibiltià di utilizzo in contesti reali ~~di accedere a tali dati~~ il forte avanzamente della computer vision e la grande quantità di video disponibili, hanno fatto dei metodi di video classification l’approccio più comune.

La **vision-based HAR**, quindi la classificazione di azioni in video clip, potrebbe essere vista come una semplice estensione naturale della classificazione di azioni in immagini ~~(quindi i singoli frame del video) con una successiva aggregazione delle predizioni.~~ ~~È evidente però come il~~ **~~progresso~~** ~~delle architettture per video classification risulti più lento e in generale~~ il task è invece più complesso ~~rispetto alla classificazione di immagini, questo~~ perché ~~la classificazione video~~ e associato **1)** costo cimputazionale maggiore ~~(es. più parametri se 2d rispetto a 2d e tempi di allenamento moltto maggiori)~~ **2)** alla necessità di avere modelli che imparino l’informazione temporale, oltre a quella spaziale di contesto ambientale, per potere distinguere azioni simili ~~(es. sedersi e alzarsi da una sedia, indistinguibili per semplice classificazione dei singoli frame)~~ **3)** e infine la complessità legata alla cattura del contesto spaziotemporale locale e globale durante scorrere dei frame (anche in quelle clip complesse che prevedano un movimento di camera oppure più azioni eseguite da più soggetti).

Il task specifico è stato quello di **Simple Action Recognition**, l’allenamento di un modello che classifichi una singola azione globale ~~(eseguita da uno o più soggetti)~~ associata a un corto video in input.

**Slide 3: Video Classification Methods [1:10, totale 2:54 ]**

I più noti metodi di classificazione video applicabili ~~al simple action recognition~~ sono suddivisibili in: metodi basati su **hand crafted features,** metodi **single stream** e metodi **two stream**. Tra i **metodi single-stream,** il metodo Single-Frame CNN, è l’implementazione più basilare del task ~~di classificazione~~ ~~(citata precedentemente)~~ che prevede l’applicazione di ~~un modello di~~ classificatore di immagini sui singoli frame per poi mediare le probabilità ottenute. Late fusion, early fusion e slow fusion sono anch’essi basati su 2D-CNN ma tengono conto della struttura temporale di sequenze di frame inserendo processi di fusione nella rete stessa. E poi ancora metodi più complessi che vedono l’utilizzo di 3D-CNN, reti LSTM e pose detection. ~~Esistono poi ancora metodi basati su 3D Convolutional Networks (C3D) modelli CNN a 3 dimensioni in grado di catturare l’informazione spaziale e temporale di movimento racchiusa in frame adiacenti. Oppure ancora mettodi che utilizzano diverse reti 2D-CNN per l’estrazione di local features e l’output viene dato in pasto a una many-to-one LSTM per l’estrazione dell’informazione temporale. Infine è possibile accoppiare un modello di pose detenction per ottenere i key point del corpo di una persona in ogni frame a una rete LSTM per determinare l’attività.~~ Gli **approcci** **two stream**, come quello quello da noi impementato basato sugli optical flow, prevedono architetture con flussi paralleli. Per quanto riguarda l’HAR su HMDB-51 la massima accuratezza ~~(come accuratezza media su 3 split)~~ attualmente raggiunta è dell’87.56% con un metodo che implementa un modello I3D. ~~(Two-Stream Inflated 3D ConvNet (I3D))~~

**Slide 3: HMDB: A Large Video Database for Human Motion Recognition [35, 3:33]**

Passando al **database**, ~~HMDB-51~~ questo contiene 51 categorie di azioni distinte, ognuna associata ad almeno 101 clip, per un totale di 6766 clip annotate, estratte principalmente da film e video di youtube. Le classi ~~di azioni~~ possono essere **raggruppate** in: 1) azioni generali del viso, come sorridere 2) azioni del viso con manipolazione di oggetti (come mangiare) 3) movimenti del corpo generali (come saltare) 4) movimenti del corpo che prevedono interazione con gli oggetti (come pettinare) e 5) infine movimenti del corpo con interazione umana come (come abbracciare)

**Slide 4: HMDB51: Data Exploration [40, 4:13] 4:40**

In aggiunta alla label ogni clip è annotata con **meta informazioni** aggiuntive, in particolare il campo 1) Visible body parts indica se la testa, la parte superiore o inferiore o l’interezza del corpo siano visibili 2) Camera motion indica se la telecamera è in movimento o statica e come è possibile osservare più di metà delle clip è movimento 3) Camera viewpoint indica se il pnto di vista della telecamera rispetto al soggetto è anteriore, posteriore a sinista o a destra 4) Number of people è il numero di pensore coinvolte e infine 5) video quality è una label su 3 livelli che indica la qualità della clip in termini di dettagli e artifatti di compressione visibili.

~~I video originali da cui sono estratte le clip variano in termini di risoluzione verticale e orizzontale e di frame rate,~~ gli autori hanno inoltr **normalizzato** le clip per favorire la consistenzascalando l’altezza di tutti i frame a 240 pixel, di conseguenza la larghezza in modo da mantenere l’aspect ratio originale e convertendo il frame rate a 30 fps.

**Slide 4: HMDB51: Data Exploration [1:10. 5:50] 5:52**

**Il processo di data preparation ~~che ha preceduto il training dei modelli~~ ha visto ~~in primo luogo~~ 1) un split del dataset in training e test set utilizzando la suddivisione suggerita dagli autori la quale garantisce che a) clip provenienti dallo stesso video non siano presenti sia nel train che nel test set b) che per ogni classe siano presenti esattamente 70 clip di train e 30 di test e che c) e che si abbia la massimizzazione del bilanciamento della proporzione relativa dei meta tag.**

**2) Segue l’estrazione dei frame e la stima degli optical flow. Con optical flow definiamo il pattern di movimento apperente di oggetti in una scena tra due frame consecutivi causato dal movimento dell’oggetto o della camera che coincide con un campo di vettori bidimensionali di spostamentto dei punti ~~tra un frame e il successivo~~. Per la stima è stato utilizzato Dual TV-L1, un algoritmo per optical flow densi (che quindi calcola il vettore di movimento per ogni pixel)~~, presente nella libreria OpenCV~~ e per ogni flow è salvata la componete orizzontale e verticale**

**3) Seguoni gli step di sampling, data augmentation, resizing, scaling e centering, le cui caratteristiche sono cambiate in base al modello implementato.**

**PARTE DI GIAN Tempi: 9:00**

**PARTE DI ALE Circa 5 minuti**

Slide 16 (1m 20 circa)

In questa slide viene mostrata l’architettura della 2 stream CNN. A sinistra la rete temporale che ha in input gli optical flow, a destra la resnet, ed in fondo ai due modelli si nota la fusione dei due output attraverso una media.

Leggi performace slides fusion. Valori simili a ResNet.

Resnet tende a fare predizione netta, ad associare una probabilità alta ad una classe candidato che è molto alta rispetto alle altre classi.

Temporal ha un’accuratezza minore della ResNet e quando i video sono poco stabilizzati tende a dare predizioni poco marcate, ad avere un buon ventaglio di classi che hanno probabilità simile o comunque alta di essere la scelte come predizione/output.

Al momento della media resnet tende a prevalere, come si può notare dai valori della fusion.

Slide 17

Preso due video appartenenti a split 1, classe 0, non utilizzata ne su train ne su test set. Entrambi con meta-label simili: camera motion, full body visibile, una left, altro front, 1 persona, qualità media.

Nonostante simili label il golf si muove solo per poco e alla fine, è quasi stabile, E il temporal funziona bene (66%).

Ride\_bike invece è sempre in movimento ed essendo tutti mossi gli optical flow la temporal fa più fatica a riconoscere azione. Come dicevo prima molte classi con buona probabilità

Label predetta comunque era giusta

Slide 18

Per il discorso delle probabilità detto prima è meglio la resnet tendenzialmente, ma quando si hanno video stabilizzati o comunque fermi anche la temporal CNN da il suo buon contributo e quindi è possibile che in questi casi performi meglio della resnet singola

Cause di bassa accuratezza:

* Come dice il paper dei creatori dataset, image degradation
* Si hanno circa 7k videos ma spesso si hanno 2-3-4 spezzoni ricavato sempre dallo stesso video. Poca diversificazione/pochi dati per trainare quindi. 1400 / 50 classi = circa 30 video per classe tra train e test
* Azione corta -> mancata durante il train (come funziona il train????????)

Modello

* Parametri di train non ottimali
* Troppi parametri, soprattutto layer densi, per pochi dati