

دانشگاه صنعتي امیرکبیر  
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش کارآموزی

موضوع کارآموزی:

**Video Swin Transformer**

|  |  |
| --- | --- |
| نام استاد کارآموزی: | دکتر صاحب الزمانی |
| نام دانشجو: | محمدرضا قادری عمودیزج |
| شماره دانشجویی: | 9627057 |

**تابستان 1399**

**تقدیر و تشکر**

بدینوسیله مراتب قدردانی و امتنان خود را خدمت،

جناب آقای مهندس حسین یعقوب زاده، سرپرست تیم توسعه دهنده پژوهشگاه توسعه فناوری‌های پیشرفته، بابت آموزش‎ها و زحمات بی‌دریغشان، تمامی کارکنان و همراهان بنده در پژوهشگاه توسعه فناوری‌های پیشرفته ،

ابراز و از تمامی زحمات آنان تشکر می‌نمایم.

محمدرضا قادری عمودیزج

شهریور 1400

چكيده

امروزه با توجه به نیازهای جامعه از نظر امنیتی و توسعه هوش مصنوعی و همینطور شناسایی رفتارهای افراد در جامعه از نظر بازدید ویدیوها و تصاویر در شبکه‌های اجتماعی و حتی استفاده از شناسایی این رفتارها در ربات‌های انسان نما برای پاسخ به حرکات پیش آمده نیاز به یادگیری و تشخیص رفتارهای تصاویر و ویدیو شده است. در اواخر قرن بیستم با توجه به رونق هوش مصنوعی از این تکنولوژی برای بررسی و تجزیه وتحلیل خودکار مبنی بر تصویربرداری استفاده شد که بینایی ماشین[[1]](#footnote-1) نامیده شد، که روز به روز در حال گسترش هست.

در این گزارش، پس از تحلیل این روش برای بررسی ویدیوها به صورتی که بتوان با بینایی ماشین به تشخیص نوع رفتاری آنها بیانجامد استفاده می‌کنیم، و نوع رفتاری و انجام تحلیل ویدیوهای، نحوه‌ی استفاده از دیتابیس و ابزارها و کتابخانه‌های لازم رو بررسی می‌کنیم.

واژه‌های کلیدی:

هوش مصنوعی، پردازش تصویر[[2]](#footnote-2)، بینایی ماشین، تشخیص رفتار ویدیو.

فهرست مطالب

[فصل 1: مقدمه](#_Toc69291746) 0

[1-1- مقدمه 0](#_Toc69291747)

[1-2- کارهای مرتبط 0](#_Toc69291748)

[1-3- معماری شبکه 0](#_Toc69291749)

[1-4- جمع‏بندی 0](#_Toc69291750)

[فصل 2: مجموعه دادگان 1](#_Toc69291746)

[2-1- مقدمه 2](#_Toc69291747)

[2-2- نحوه دسترسی 2](#_Toc69291748)

[2-3- جزئیات مجموعه دادگان 2](#_Toc69291749)

[2-4- جمع‏بندی 2](#_Toc69291750)

[فصل 3: راه‏اندازی کد به صورت برخط (Online) 3](#_Toc69291751)

[3-1- مقدمه 4](#_Toc69291752)

[3-2- نصب پکیج‏ها و پیش‏نیازها 4](#_Toc69291753)

[3-3- نحوه دسترسی 4](#_Toc69291754)

[3-4- راه‏اندازی سریع 4](#_Toc69291755)

[3-5- بررسی جزئیات 5](#_Toc69291756)

[3-6- جمع‏بندی 6](#_Toc69291757)

[فصل 4: راه‏اندازی کد به صورت برون خط (Offline) 7](#_Toc69291758)

[4-1- مقدمه 8](#_Toc69291759)

[4-2- نحوه دسترسی 8](#_Toc69291760)

[4-3- نصب پیش‏نیازها به صورت Offline 8](#_Toc69291761)

[4-4- راه‏اندازی سریع 8](#_Toc69291762)

[4-5- بررسی جزئیات کد 8](#_Toc69291763)

[4-6- جمع‏بندی 9](#_Toc69291764)

[فصل 5: جمع‏بندی و پیشنهادات 12](#_Toc69291765)

[5-1- مقدمه 13](#_Toc69291766)

[5-2- جمع‏بندی کلی و پیشنهادات 13](#_Toc69291767)

[5-3- کارهای آینده 13](#_Toc69291768)

[5-4- جمع‏بندی 13](#_Toc69291769)

[مراجع 14](#_Toc69291770)

فهرست اشکال

[شکل (1-1) تصویر نمونه 2](#_Toc34822082)

**فهرست جداول**

[جدول (1-1) جدول نمونه 2](#_Toc34821840)

1. مقدمه
   1. مقدمه
   2. بینایی ماشین و پردازش تصویر
   3. ترانسفورماتور چرخش تصویری[[3]](#footnote-3)
   4. جمع بندی
2. معرفی محل کارآموزی
3. مجموعه دادگان
   1. مقدمه

در این پروژه با توجه به این که ما دنبال این هستیم که حرکات انسان رو بشناسیم و یاد بگیریم نوع حرکات انسان در ویدیو ها، از 3 نوع دیتاست استفاده می‌کنیم که از نظر نوع داده‌های آموزشی و تستی متفاوت هستند. (مقدار و نوع پخش شدن داده‌ها)

اولین دیتاستی که در مورد آن خواهیم گفت Kinetic-400 ، سپس Kinetic-600 و در نهایت  
 Something-Something v2 رو داریم.

به طور کلی Kineticدارای یک مجموعه از بیش از 650000 ویدیو با کیفیت و اندازه مناسب است که کلاس بندی شده اند با توجه به ورژن آن دیتاست هست و این ویدئوها شامل فعل و انفعالات بین انسان و شیء مانند نواختن سازها ، و همچنین تعاملات انسان با انسان مانند دست دادن و در آغوش گرفتن است. که هر کدام از کلاس ها حداقل به اندازه تعداد کلاس ها ویدیو دارد، که هر ویدیو فقط یک حرکت و حدودا 10 ثانیه زمان دارد. اکثر این ویدیو ها از یوتیوب جمع آوری شده است.

Something-something v2 هم ورژنی آپدیت شده است که مجموعه بزرگی از کلیپ های ویدیویی دارای برچسب که نشان دهنده اقدامات اولیه پیش فرض انسان با اشیا روزمره را داراست که با استفاده از جمعیت زیادی از کارگران جمع آوری شده است. این به مدلهای یادگیری ماشین اجازه می دهد تا درک دقیقی از اقدامات اساسی که در دنیای فیزیکی رخ می دهد توسعه دهند.

* 1. نحوه دسترسی

ویدیوی‎های دیتاست‌ها در دوحالت قابل دسترس خواهند بود اولی به صورت آنلاین و دیگری به صورت آفلاین(دانلود شده). ما با حالت اول کار داریم در این حالت ما یک فایل json یا csv داریم که دارای مشخصاتی به مانند label,youtube\_id,time\_start,time\_end,split که Split به معنی نوع استفاده‌ای برای ویدیو است(یادگیری/تست/اعتبارسنجی). این ویدیوها در فایلهای مانند csv به صورت زیر برای مثال ذخیره می‌شود.

{

"---QUuC4vJs": {

"annotations": {

"label": "testifying",

"segment": [

84.0,

94.0

]

},

"duration": 10.0,

"subset": "validate",

"url": "https://www.youtube.com/watch?v=---QUuC4vJs"

},

"--GkrdYZ9Tc": {

"annotations": {

"label": "washing feet",

"segment": [

0.0,

10.0

]

},

"duration": 10.0,

"subset": "validate",

"url": "https://www.youtube.com/watch?v=--GkrdYZ9Tc"

},

.

.

.

.

}

با داشتن فایل بالا می‌توان با خواند اطلاعات مورد نیاز از این فایل و دسته بندی قسمت های مختلف برای کلیپ ها (به عنوان مثال به صورت یک دیکشنری) در می‌آوریم و شروع به کار می‌کنیم.

* 1. جزئیات مجموعه دادگان

[Kinetic-400:](https://arxiv.org/pdf/1705.06950.pdf)  در حال حاضر در این ورژن از دیتاست 306245 ویدیو موجود هست که هر کلاس 400 - 1150 تا ویدیو دارد که این ویدیو ها به سه قسمت تقسیم می‌شوند که یک قسمت برای یادگیری[[4]](#footnote-4) که برای هر کلاس بین 250 - 1000 ویدیو دارد ، 50 تا برای اعتبارسنجی[[5]](#footnote-5) و در نهایت 100 ویدیو برای تست[[6]](#footnote-6) برای هر کلاس موجود است.



تعداد کلیپ ها برای هر کلاس در پارتیشن های یادگیری/اعتبارسنجی/تست

در مقاله ارجاع داده شده درباره این دیتاست درباره موارد مختلفی مانند گیج کننده ترین کلاس ها و سخترین کلاس ها برای تصمیم گیری صحبت شده.(در مورد 400 کلاس که چه موضوعی هست گفته شده)

[Kinetic-600:](https://arxiv.org/pdf/1808.01340v1.pdf) در حال حاضر در این ورژن از دیتاست 495547 ویدیو موجود هست که هر کلاس 600 - 1200 تا ویدیو دارد که این ویدیو ها به سه قسمت تقسیم می‌شوند که یک قسمت برای یادگیری که برای هر کلاس بین 450 - 1000 ویدیو دارد ، 50 تا برای اعتبارسنجی و در نهایت 100 ویدیو برای تست و یک مجموعه تست اضافی حدودا 50تایی برای فعالیت های خاص در این ورژن در نظر میگیریم برای هر کلاس موجود است. نسبت به kinetic-400 از 368 تا از کلاس‌های آن استفاده شده است، برای بقیه 32 کلاس به دلیل اینکه تقاطع زیادی با کلاس های دیگر داشت و شاید برای تشخیص دچار مشکل میشدیم تغییر کرد.



تعداد کلیپ ها برای هر کلاس در پارتیشن های یادگیری/اعتبارسنجی/تست (مقایسه با 400)

در مقاله ارجاع داده شده در بالا درباره این دیتاست درباره موارد مختلفی مانند گیج کننده ترین کلاس ها و سخترین کلاس ها برای تصمیم گیری صحبت شده.(کلاس های اضافه شده نسبت به 400 در انتهای مقاله ارجاع داده شده صحبت شده)



شکل 1- نمونه‌ای از ویدیو‌های Kinetic

[Something-Something v2:](https://20bn.com/datasets/something-something) در حال حاضر در این دیتاست 220847 ویدیو موجود هست که شامل 168913 تا ویدیو برای یادگیری و 24777 ویدیو برای اعتبارسنجی و 27175 ویدیو نیز برای تست در نظر گرفته شده با 174 برچسب متفاوت. در این دیتاست بر مبنای اسم برخورد می‌شود برای مثال قراردادن چیزی بر روی چیز دیگر. در ورژن جدید تر این دیتاست TwentyBN 30408 تا شی متمایز از یکدیگر داریم.

* 1. جمع بندی

1. راه اندازی کد
   1. مقدمه

برای هر پروژه‌ای از تعدادی از نرم افزارها و کتابخانه های زبان های نوشته شده کدها برای اجرا و راه اندازی استفاده می کنیم. در این پروژه با توجه به اینکه از زبان پایتون استفاده شده است بایستی یکسری از کتابخانه‌های مربوط به این زبان و یکسری ابزار مربوط به این زبان در راستای کار کردن با ویدیو را بر روی سیستم خود نصب کنیم.

* 1. نصب پکیج‏ها و پیش‏نیازها

با توجه به اینکه زبان مورد استفاده در این پروژه پایتون هست بهتر است در ابتدا از نصب شدن زبان مورد نظر اطمینان کافی داشته باشیم و حدالمقدور آخرین نسخه آن در اختیار باشد.



شکل 1-3: نصب و نسخه پایتون

Pytorch یک کتابخانه‌ای از زبان پایتون هست، که منبع باز برای یادگیری ماشین است که بر مبنای کتابخانه Torch هست که در کاربردهای بینایی ماشین و پردازش زبان استفاده می‌شود که توسط گروه هوش مصنوعی فیسبوک طراحی شده است. برای نصب آخرین ورژن این کتابخانه می توانیم از دستور Pip install torch استفاده کنیم.



شکل 2-3: نصب pytorch

برای چک کردن ورژن نسخه Pytorch میتوانیم از کامپایل فایل pytorch-V.py موجود در پروژه استفاده کرد.

CUDA یک نرم افزار (پلتفرم) محاسباتی موازی و مدل برنامه نویسی است که توسط Nvidia برای محاسبات عمومی در GPU ها توسعه یافته است. CUDA توسعه دهندگان را قادر می سازد تا با استفاده از قدرت پردازنده های گرافیکی در بخش موازی سازی محاسبات ، برنامه های کاربردی با سرعت بالا را افزایش دهند.

برای نصب آن بایستی فایل نصبی آن از آدرس [cuda-download](https://developer.nvidia.com/cuda-downloads) دریافت کرده و نصب کنیم (البته نیاز به نصب visual studio نیز هست).



شکل 2-3: نسخه cuda

مجموعه کامپایلر GNU یک کامپایلر بهینه ساز است که توسط پروژه GNU تولید شده و از زبانهای برنامه نویسی مختلف ، معماری سخت افزار و سیستم عامل پشتیبانی می کند. بنیاد نرم افزار آزاد GCC را به عنوان نرم افزار رایگان تحت مجوز عمومی GNU توزیع می کند. برای نصب آن از Cygwin استفاده میکنیم تا پکیج های لازم را برای آن بالا بیاورد.



شکل 2-3: نسخه gcc

Mmcv یک کتابخانه اساسی و کاربردی پایتون برای بینایی ماشین است. برای نصب آن از

* 1. نحوه دسترسی

متن.

* 1. راه‏اندازی سریع

متن.

نمونه کد:

args = [

        '--name', \_\_file\_\_.split('/')[-1].split('.')[0],  # name is filename

        '--print-freq', '1',

        '--data', 'D:/Charades/Charades\_v1\_rgb/Charades\_v1\_rgb',

        '--dataset', 'charadesrgb',

        '--cache-dir','./cache/',

        '--pretrained',

        #'--evaluate',

    ]

نمونه خطا:

|  |  |
| --- | --- |
| خطای 1: خطای عدم وجود کتابخانه | نحوه رفع این خطا |
|  |  |

* 1. بررسی جزئیات

متن.

.

* 1. جمع‏بندی

متن.

1. جمع بندی و پیشنهادات
   1. مقدمه

متن.

* 1. جمع‏بندی کلی و پیشنهادات

متن.

* 1. کارهای آینده

متن.

* 1. جمع‏بندی

متن.

مراجع

[1] Q. Liu, A. H. Sung, and M. Qiao, "Video steganalysis based on the expanded markov and joint distribution on the transform domains detecting msu stegovideo," in *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, 2008: IEEE, pp. 671-674.

1. Machine vision [↑](#footnote-ref-1)
2. Image Processing [↑](#footnote-ref-2)
3. Video Swin Transformer [↑](#footnote-ref-3)
4. Training [↑](#footnote-ref-4)
5. Validations [↑](#footnote-ref-5)
6. Test [↑](#footnote-ref-6)