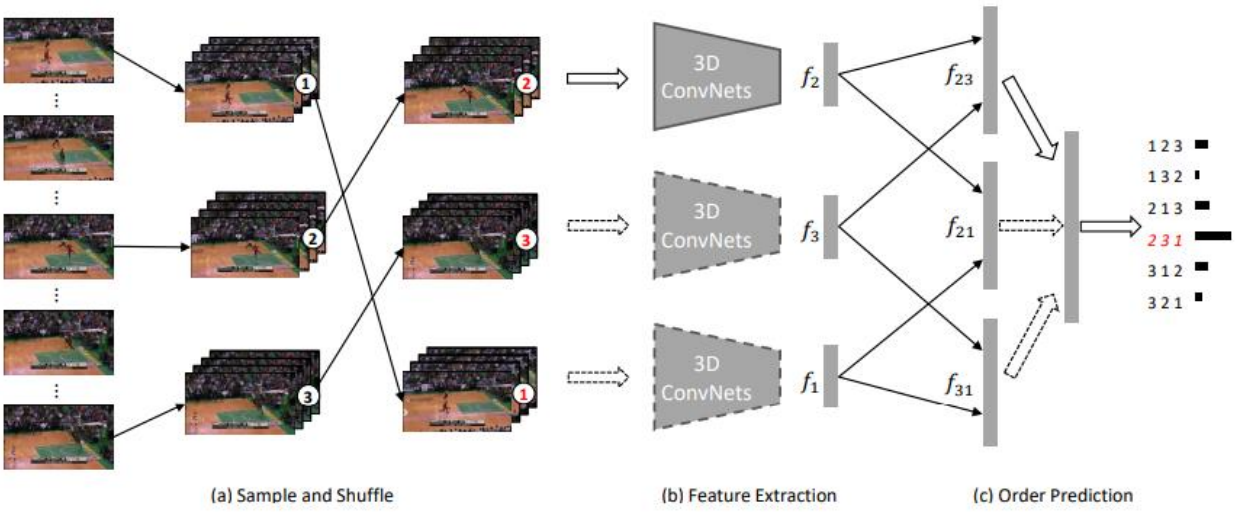
به نام خدا

در این مقاله یک نوع pretext task برای ویدیو پیشنهاد شده است که در آن برش هایی(clip) از ویدیو را به عنوان ورودی میگیرد و ترتیب این برش ها را پیش بینی میکند. مدلی که در اینجا استفاده میشود unsupervised است. وزن های آموزش دیده شده این مدل میتواند به عنوان وزن های پیش آموخته در مدل های دیگر برای کار هایی مانند action recognition استفاده و fine tune شود.

توضیح روش:



1- Sample and shuffle:

برش هایی که با هم همپوشانی ندارند نمونه برداری میشوند و با یک ترتیب تصادفی بُر میخورند.

2- Feature extraction:

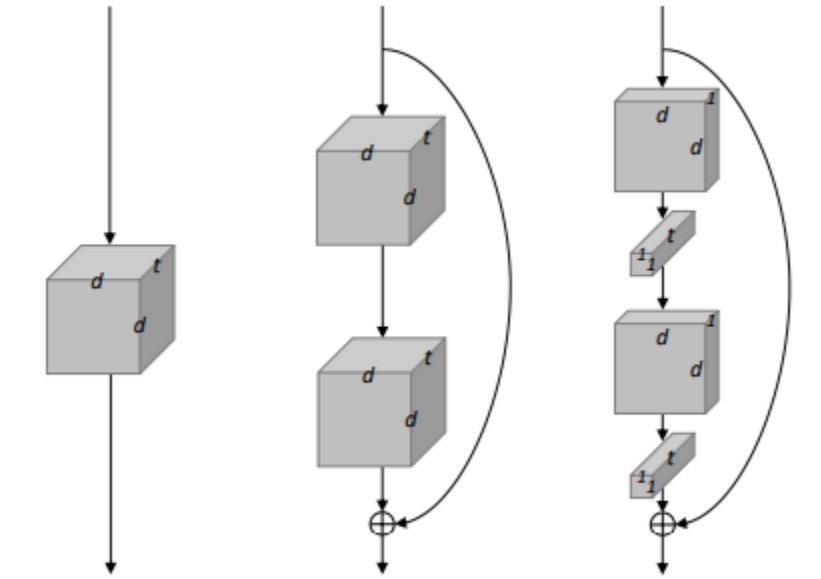
یک شبکه سه بعدی کانولوشنی برای استخراج ویژگی ها استفاده میشود. این شبکه در هیچ دیتاستی از قبل آموزش ندیده است که وزن های اولیه داشته باشد.

3- Order prediction:

ویژگی های استخراج شد به صورت جفت جفت concatenate میشوند و به لایه های کاملا متصل که برای پیش بینی ترتیب هستند وصل میشوند.

این فریمورک میتواند به صورت end to end آموزش داده شود و convent سه بعدی آن به عنوان استخراج کننده ویژگی ویدیو یا به عنوان وزن های پیش آموخته استفاده شود.

در این مقاله از سه نوع مختلف cnn سه بعدی به عنوان استخراج کننده ویژگی استفاده شده.



1- C3D conv blocks:

از فیلتر کانولوشنی سه بعدی کلاسیک با سایز t × d × d استفاده شده که روی هم انباشته میشوند تا شبکه C3D را تشکیل دهند.

2- R3D conv blocks:

کرنل کانولوشنی سه بعدی کلاسیک با اتصال shortcut.

3- R(2 + 1)D conv blocks:

کرنل سه بعدی ای که به دو قسمت تجزیه شده. کرنل دو بعدی مکانی با سایز 1 × d × d و کرنل یک بعدی زمانی با سایز t × 1 × 1.

نتایج:

3D CNNs	C3D	R3D	R(2+1)D
Accuracy	68.5	68.4	64.2

Table 1. **Clip order prediction results on UCF101**. C3D, R3D and R(2+1)D networks are trained with clip order prediction framework separately.

Methods	Top1	Top5	Top10	Top20	Top50
Jigsaw [27]	19.7	28.5	33.5	40.0	49.4
OPN [22]	19.9	28.7	34.0	40.6	51.6
Büchler et al. [1]	*25.7*	*36.2*	*42.2*	*49.2*	*59.5*
C3D (random)	16.0	22.5	26.7	31.4	39.3
C3D	12.5	29.0	39.0	50.6	**66.9**
R3D (random)	10.5	17.2	21.5	27.0	36.7
R3D	14.1	**30.3**	**40.0**	**51.1**	66.5
R(2+1)D (random)	10.2	17.3	21.9	27.7	38.5
R(2+1)D	10.7	25.9	35.4	47.3	63.9

Table 2. **Frame and clip retrieval results on UCF101.** The methods in top row are based on 2D CNNs while 3D CNNs in our framework are presented in bottom row.

Methods	Top1	Top5	Top10	Top20	Top50
C3D (random)	7.7	12.5	17.3	24.1	37.8
C3D	7.4	22.6	**34.4**	48.5	**70.1**
R3D (random)	5.5	11.3	16.5	23.8	37.2
R3D	7.6	**22.9**	**34.4**	**48.8**	68.9
R(2+1)D (random)	4.6	11.1	16.3	23.9	39.3
R(2+1)D	5.7	19.5	30.7	45.8	67.0

Table 3. **Clip retrieval results on HMDB51.** The 3D CNNs used here are self-supervised trained on split 1 of UCF101 merely.

همانطور که مشاهده میشود این روش با سه نوع کانولوشن سه بعدی که معرفی شد نسبت به بقیه روش ها بهبود داشته اند.