Transfer Learning

Sign Language Video Classification

هدفي

اليوزر يرفع فيديو

البرنامج يشوف الفيديو

يقرأ الإشارة اللي فيه

(go – banana – talk …) الموديل يطلع **الكلمة**

Mediapipe Landmarks

:الفكرة\

بدل ما تعلّم الموديل يشوف الفيديو كله (تقيل جدًا)…
. نخلّيه يشوف **نقاط اليد + الجسم** فقط

:المطلوب Preprocessing\

1) استخراج فريمات من الفيديو

- مش لازم عدد معين
- بتاخد فريم كل X ms
- (مثلاً 30 فريم من الفيديو كله)

2) Mediapipe Pose/Hands

- تستخرج:

- نقطة لليد 21
- أو 33 نقطة للجسم
- أو الاثنين = أقوى حاجة

3) Normalize للإحداثيات

- توحيد القيم بين 0 و 1
- أو Normalize relative to shoulder

4) Pad / Sampling

علشان كل الفيديوهات يبقى عندها نفس عدد الفريمات
مثال :

```
[30 frames × 21 points × 3 coords]
```

5) Label Encoding

تحويل الكلمات لأرقام.

6) Train / Val / Test split

augmentation على val/test ومتحطش.

شكل الـ Input للموديل:

```
(batch, num_frames, num_points*coords)
```

مثال :

```
(1, 30, 21*3)
```

الموديل المناسب:

- LSTM
- GRU
- Transformer Encoder صغير
- أو Temporal CNN

مميزات:

- أسرع ب 10 مرات
- GPU مش محتاج
- دقة عالية للإشارات
- ممتاز لأي فيديو من أي كاميرا

جاهزة Video Models من Transfer Learning)

الفكرة:

ندخل الفيديو نفسه، مش النقاط.

المطلوب Preprocessing:

1) Extract fixed number of frames

ضروري جدًا
مثال :

```
16 frame per video
```

2) Resize كل frame

إلى:

```
224×224
```

(زي كل الموديلات الكبيرة)

3) Normalize

حسب المتدرب عليه الموديل الأصلي:

```
mean = [0.485, 0.456, 0.406]
```

```
std  = [0.229, 0.224, 0.225]
```

4) ترتيب البيانات:

```
(batch, num_frames, 3, 224, 224)
```

5) Label Encoding

زي الأولى بالظبط.

6) Split (Train/Val/Test)

7) Augmentation فقط لـ Train

- flip
- random crop
- brightness jitter
- temporal jitter

الموديلات المناسبة:

- I3D
- R(2+1)D
- SlowFast
- VideoMAE
- TimeSformer
- ViViT

مميزات:

- دقة عالية
- موديلات قوية ومعروفة عالميًا

عيوب:

- تقيل جدًا جدًا
- محتاج GPU قوي
- preprocessing كبير
- مش عملي للتطبيقات الحقيقية