Master in Computer Vision
Barcelona

UAB Universitat Autònoma de Barcelona
Universitat de Barcelona
UOC Universitat Oberta de Catalunya
UPC Universitat Politècnica de Catalunya BarcelonaTech
upf. Universitat Pompeu Fabra Barcelona

Week 7:
MULTI-MODALITY

MCV – C6

Team 3:
Iker García Fernández
Georg Herodes
Pablo Vega Gallego
Sígrid Vila Bagaria

0. Contents

1. **Week 5** - Multi-view inference I
2. **Week 6** - Multi-view inference II
3. **Week 7** - Multimodality
    a. Alternative modality
    b. Multimodal approach
4. **Conclusions**

A crab doing pull-ups

Week 5

Multi-view inference I

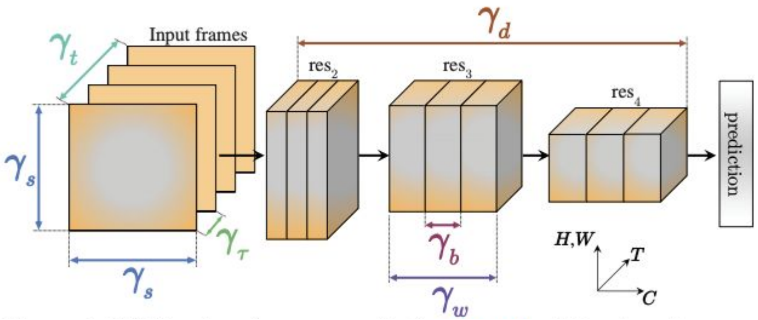
1. Week 5: Multi-view inference I - Task 1 & 2 Baseline

Initial conditions:

We use the default parameters of the model (**X3D-XS**):

- **Crop Size:** 128
- **Temporal Stride:** 12
- **Clip length:** 4
- **Batch Size:** 16
- ***Patience:* *3***

Also, we added the **Early Stopping.**



Results:

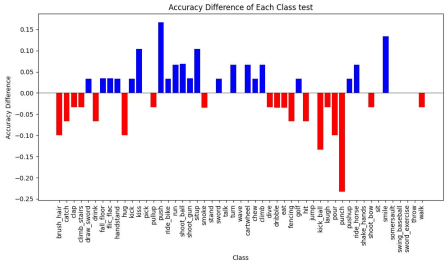
Test accuracy is: 0.17996
Train accuracy is: 0.35897

1. **Week 5:** Multi-view inference I - Task 3 Inference

Temporal Inference - parameter search:

- **Clip Length:** [**4**, 8, 16]
- **Crop Size:** [150, 182, 200, **250**]
- **N_t:** [**1**, 2, 4, 8, 16]
- **Temporal stride:** [4, 8, **12**, 16]

Config parameter	Importance	Correlation
clip_length		
nt		

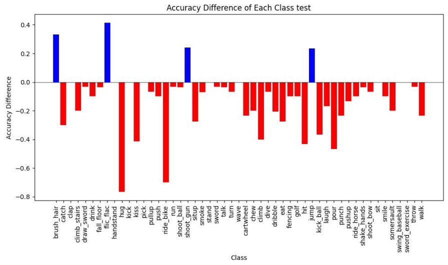


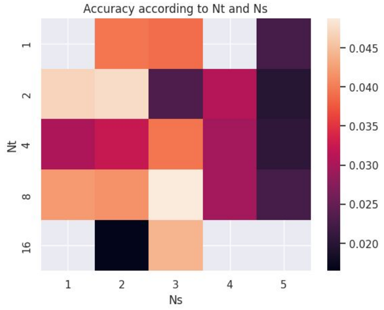
$$\text{Overlap \%} = \frac{\text{Total Length of Clips}}{\text{Length of Overlap}} \times 100\%$$

Spatio-Temporal Inference - parameter search:

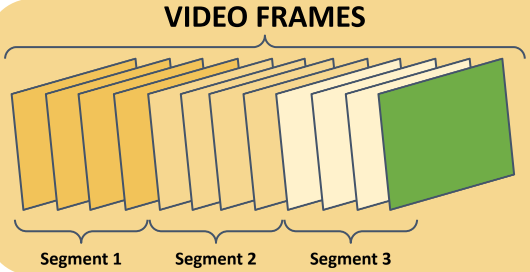
- **Clip Length:** [**4**, 8, 16]
- **Crop Size:** [150, **182**, 200, 250]
- **N_t:** [1, 2, 4, **8**, 16]
- **Temporal stride:** [4, 8, 12, **16**]
- **N_s:** [1, 2, **3**, 4, 8, 16]

Config parameter	Importance	Correlation
clip_length		
temporal_stride		



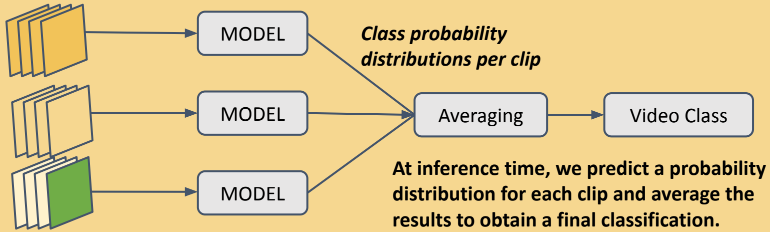


1. Week 5: Multi-view inference I - Task 4 TSN Training



Split each video into N equally sized segments.

The last frame is duplicated to create equally sized clip tensors.



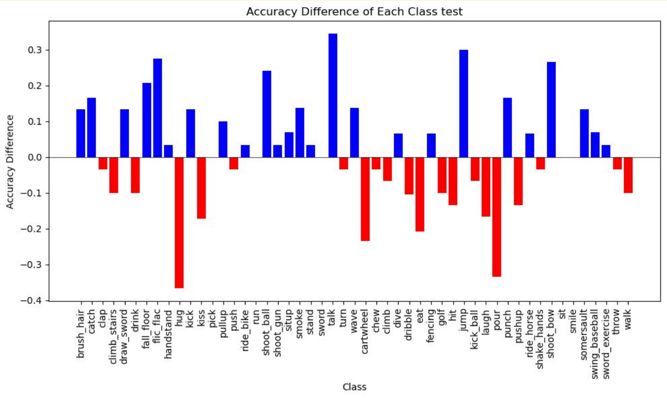
At inference time, we predict a probability distribution for each clip and average the results to obtain a final classification.

Results:

Test accuracy 0.186

LR	Test accuracy	Train accuracy
1e-4 (BL)	**0.176**	**0.342**
5e-5	0.163	0.259
1e-5	0.11	0.1275

N. of segments	Test accuracy	Train accuracy
2	0.164	0.335
3 (BL)	0.176	0.342
5	**0.186**	**0.373**



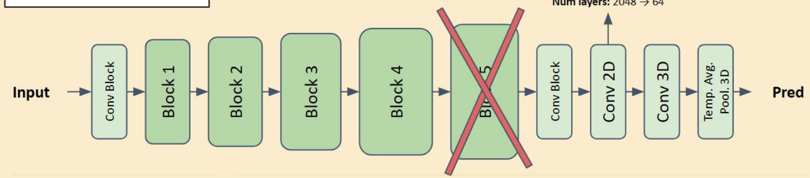
Week 6

Multi-view inference II

2. Week 6: Multi-view inference II - Task 1 Changing the model

Small Model

MoViNet-A0 architecture

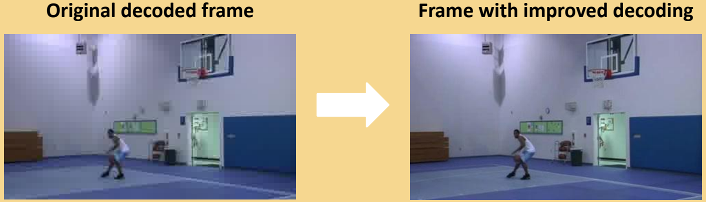


Hyperparameters:

- **Batch size:** 16
- **Crop size:** 182
- **Clip length:** 4
- **Temporal stride:** 12
- **Optimizer:** Adam
- **Loss:** CrossEntropy
- **LR:** 1e-4
- **Epochs:** 50
- **Pretrained:** True
- **TSN:** No
- **Multi-Clip testing:** 1x1

	Params (M)	FLOPs (G)	Train Acc. (%)	Valid. Acc. (%)	Test Acc. (%)
BL	0.31	0.09	39.02	16.53	19.25
BEST	0.06	0.02	**57.25**	**37.71**	**41.79**

Big Model



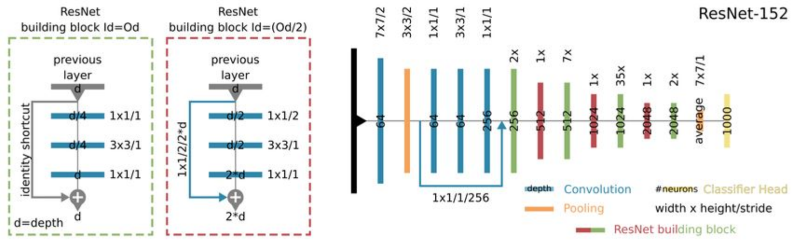
Hyperparameters:

- **Model:** X3D-M
- **Batch size:** 16
- **Crop size:** 256
- **Clip length:** 16
- **Temp. stride:** 5
- **TSN:** No
- **Loss:** CrossEntropy
- **LR:** 1e-3
- **Multi-Clip testing:** 1x1
- **Data augm.:**
  - RandomResizedCrop
  - RandomHorizCrop
  - ColorJitter

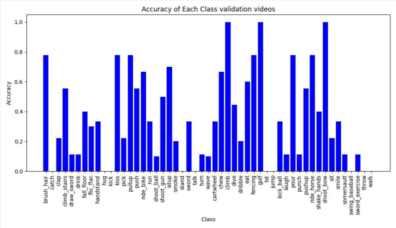
	Params (M)	FLOPs (G)	Train Acc. (%)	Valid. Acc. (%)	Test Acc. (%)
BL	0.31	0.67	39.2	22.0	24.2
BEST	0.31	0.67	**99.8**	**69.3**	**71.8**

2. Week 6: Multi-view inference II - Task 2 Temporal dynamics

Predicting Frame Class

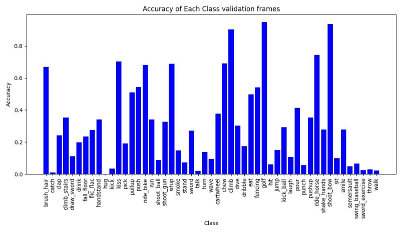


By video:



Val acc by video: 0.3516

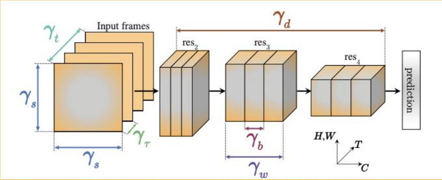
By frame:



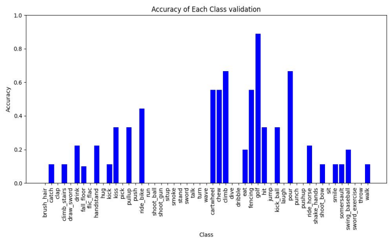
Val acc by frame: 0.3519

Shuffling Frames

Baseline Model (X3D-xs)

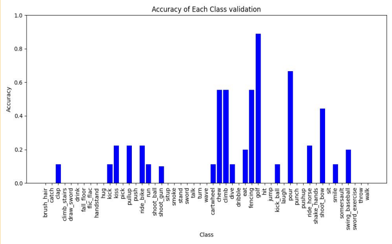


Ordered frames:



Val acc by video: 0.1892

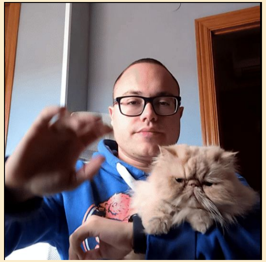
Shuffled frames:



Val acc by frame: 0.1292

2. Week 6: Multi-view inference II - Experiments

Some tests we did with videos of us:



No temporal information
Max class → laugh
Wave: 0.0412

S: wave: 0.33
B: wave: 1.00

No temporal information
Max class → laugh
Wave: 0.0129

S: laugh: 0.99
B: wave: 0.95

No temporal information
Max class → drink
Pour: 0.2871

S: talk: 0.43
B: pour: 0.55
drink: 0.38

No temporal information
Max class → smile
Smile: 1.0

S: wave: 0.41
B: brush_hair: 1.00

No temporal information
Max class → brush_hair
Brush_hair: 0.6777

S: brush_hair: 0.99
B: brush_hair: 1.00

No temporal information
Max class → pour
Chew: 0.3653

S: wave: 0.32
B: wave: 1.00

Week 7

Multimodality

3. Week 7: Multimodal - Alternative modality

OF pre-extraction

GMFlow [1]

WHY OPTICAL FLOW?

3 main reasons:

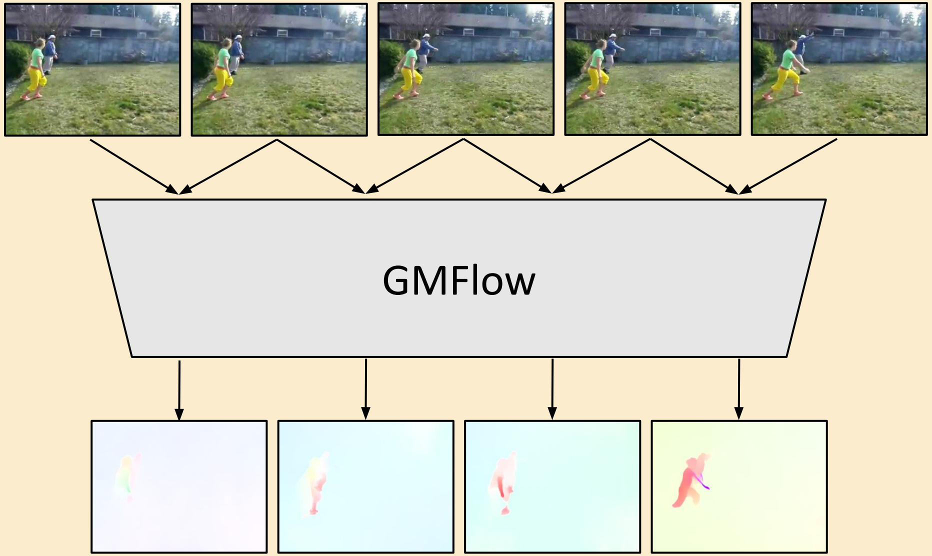
- Motion Capture
- Robust Feature Representation
- Handling Complex Scenes

	Device	# Images	Inf. time	Total time
Pre-extraction		634258		

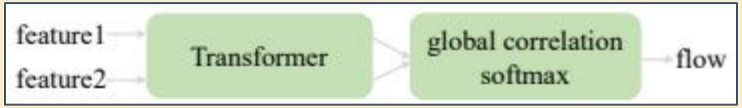
[1] Haofei Xu1, Jing Zhang2 et al. *"GMFlow: Learning Optical Flow via Global Matching"*. arXiv preprint arXiv:2111.13680v4 17 Jul 2022

3. Week 7: Multimodal - Optical Flow

OF pre-extraction



GMFlow [1]



- **Global matching**
- **Robustness to Appearance Changes**
- **High Accuracy**

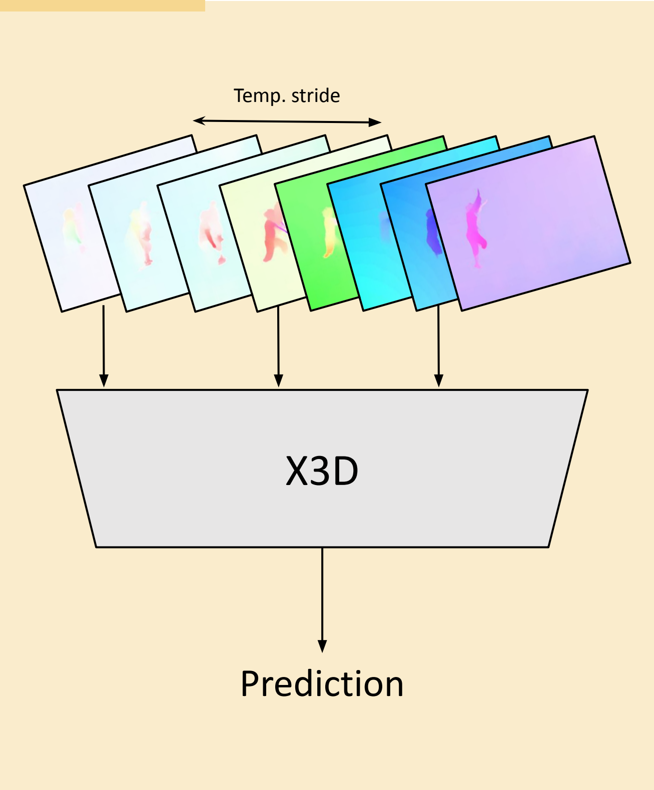
Times

	Device	# Images	Inf. time	Total time
Pre-extraction	RTX 3090	634258	0.035s/image	6h 10min

[1] Haofei Xu1, Jing Zhang2 et al. *"GMFlow: Learning Optical Flow via Global Matching".* arXiv preprint arXiv:2111.13680v4 17 Jul 2022

3. Week 7: Multimodal - Optical Flow

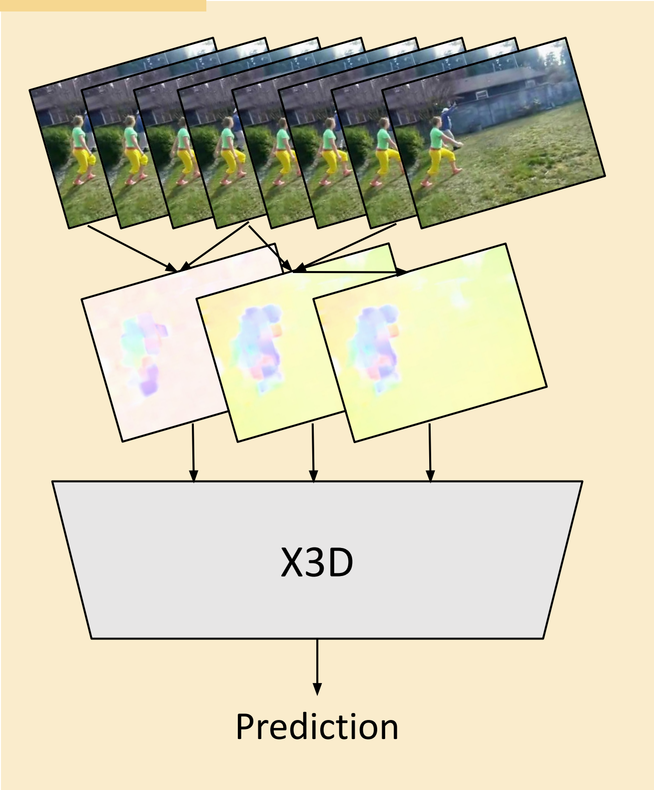
Method 1



- The pre-extracted OF estimations are used the same way frames were used in the original implementation.

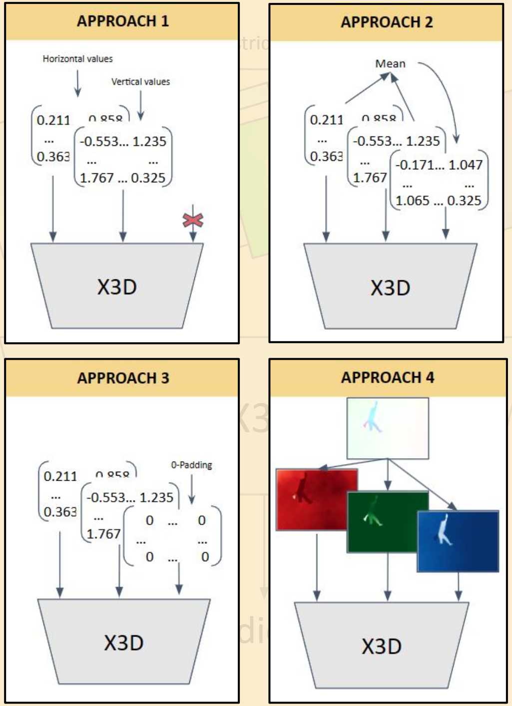
- The OF is computed online:
  - **FarneBack** algorithm is used as it is faster.

Method 2



3. Week 7: Multimodal - Optical Flow

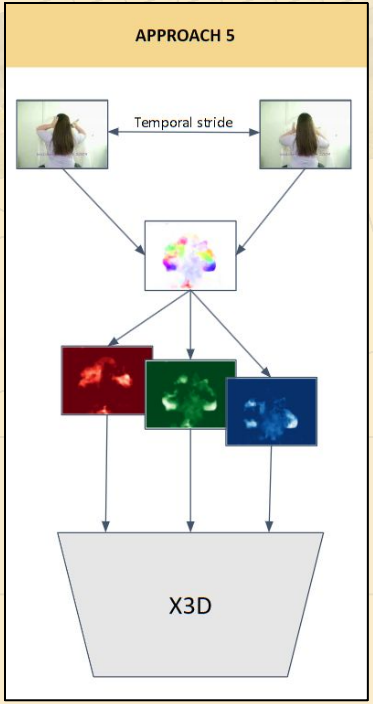
Method 1



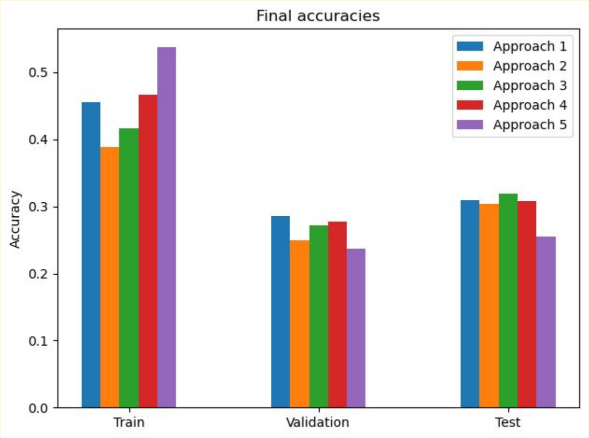
- **Approach 1:** Use OF values and adapt the net to use 2 input channels.
- **Approach 2:** Fill the 3rd channel with the mean of the vert. and horiz. OF.
- **Approach 3:** 0-pad the 3rd channel.
- **Approach 4:** Convert the OF values into an RGB representation and use it.

- **Approach 5:** Use RGB visualizations, as in approach 4.

Method 2



3. Week 7: Multimodal - Optical Flow



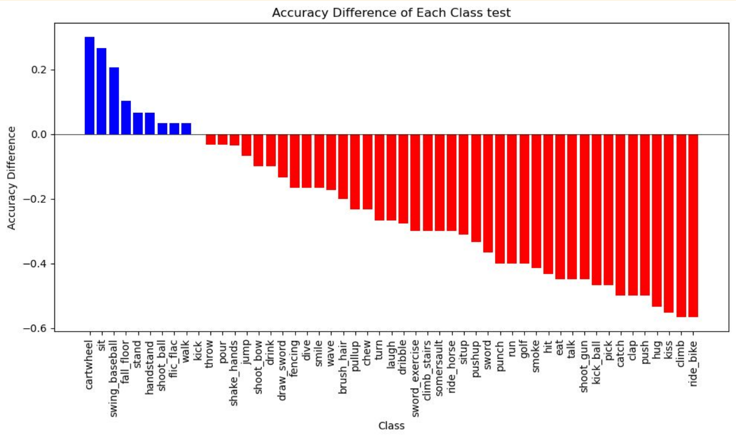
- The similar results of the first 4 approaches shows that the model is able to extract the same information from the different inputs.

- Computing the optical flow between distant frames has produced noisy estimations which led to worse results.

Best Approach:
0-padding 3rd Channel

	RGB	FLOW
# Params (M)	3.1	3.1
GFLOPs	0.9	0.9
Train Acc. (%)	99.8	85.9
Val. Acc. (%)	69.3	32.2
Test Acc. (%)	71.8	38.1

Hyperparameters:
- **Model:** X3D-XS
- **Clip length:** 16
- **Crop size:** 256
- **Batch size:** 16
- **Optimizer:** ADAM
- **LR:** 1e-4
- **Temporal stride:** 8

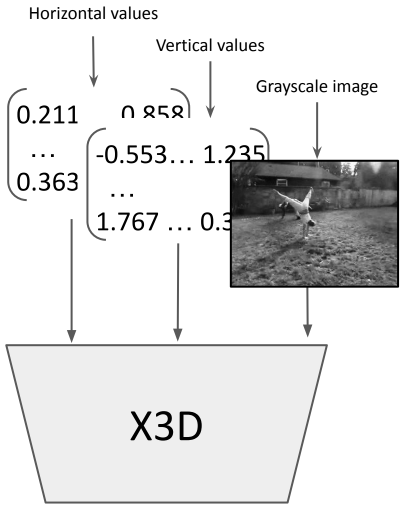


Conclusions:
Optical flow data with X3D performs worse than RGB.
This finding is perfectly expected, OF data encodes:
a) different information than RGB images
b) information in a different format than RGB images
X3D still somewhat successfully performs action classification.

3. Week 7: Multimodal - Optical Flow

3. Week 7: Multimodal - Early fusion

Early Fusion Approach



First Early Fusion Test:

Combine 2D Optical Flow data with greyscale frames:

- 3D data to use with X3D
- Combines OF and visual modalities

Test accuracy increases by 18%.

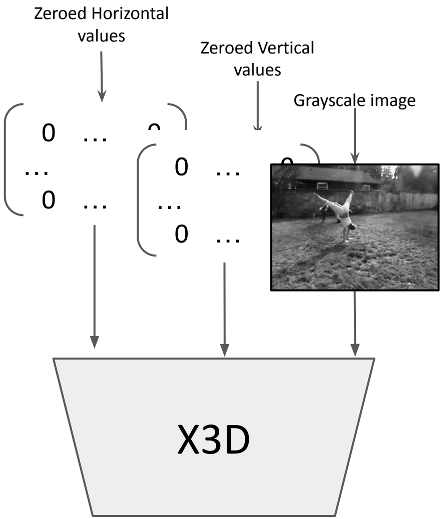
Ablation study:

Evaluate the contribution of optical flow modality to this task.

Adding OF data in this way actually made the model perform **worse**.

Approach	Train Acc	Val Acc	Test Acc
Early Fusion (Flow + Gray)	0.995	0.519	0.564
Ablation Study (Zeroes + Gray)	0.998	0.532	**0.587**

Ablation Study



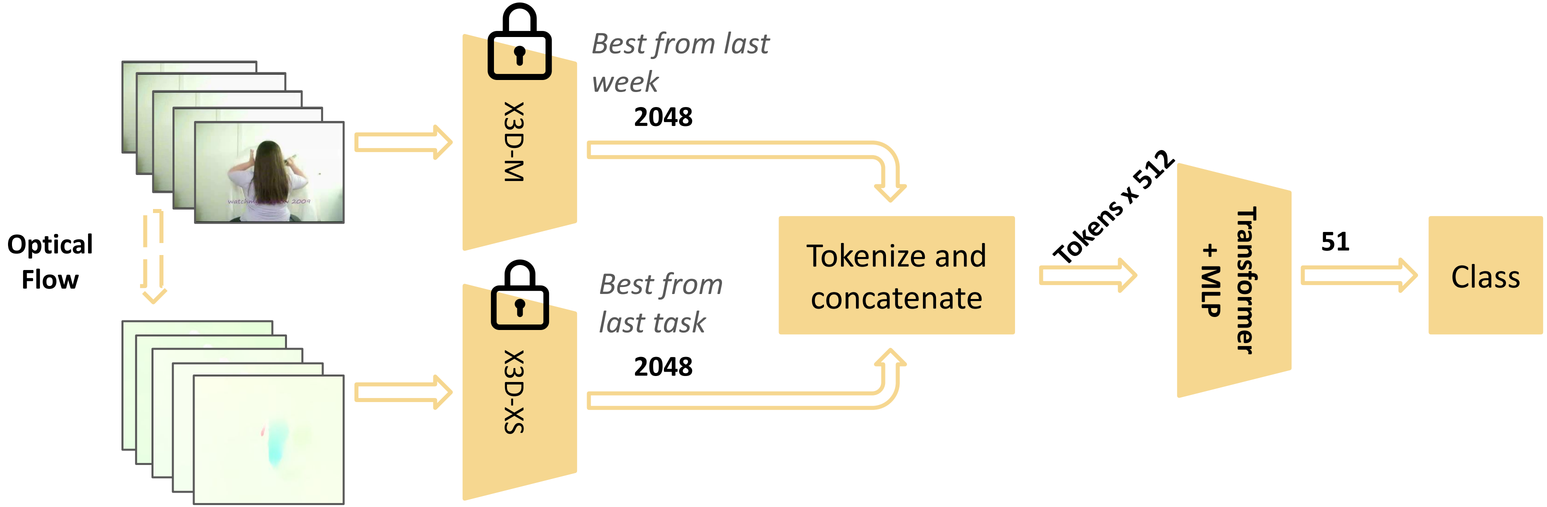
3. Week 7: Multimodal - Early fusion

Early fusion pipeline

1. Tokenize the **last layer** (before logits) of each of the previously trained models.
2. Concatenate the tokens.
3. Transformer + MLP to fusion the predictions.
4. Predict the final class.

Motivation

- Transformer very suitable because of the **self-attention**.
- We think (and hope!) that early fusion will work better than late fusion.
- Enables better integration of complementary features from the start.



3. Week 7: Multimodal - Early fusion

Hyperparameter search:

- **Epochs:** 20, 30, 40, **50**.
- **Optimizer: Adam**, SGD.
- **LR:** 1e-5, 5e-5, **1e-4**, 5e-4, 1e-3.
- **Batch size:** 4, 8, **16**.

Weights & Biases

- Overfitting.
- Improvement regarding baseline.
- More than 1 day to train → 1d 35m 15s.

Train acc: 0.3589

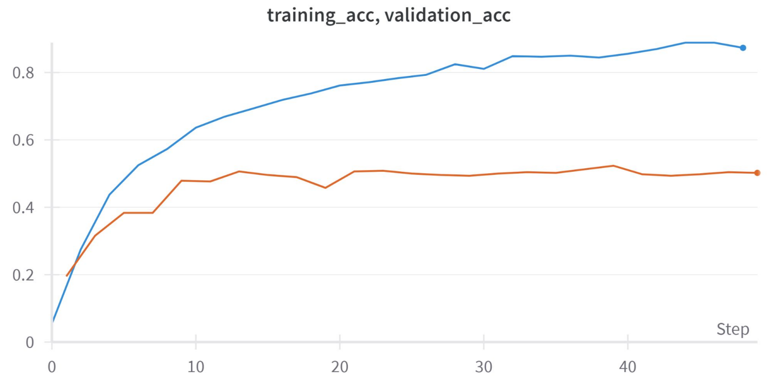
Val acc: 0.1779

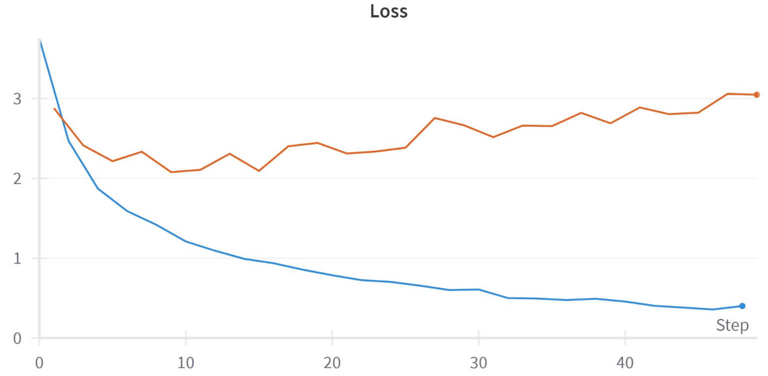
Baseline

Train acc: 0.8636

Validation acc: 0.5543

Early Fusion

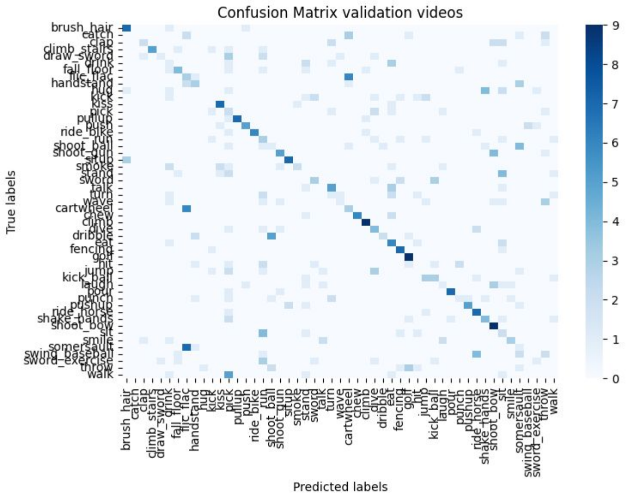


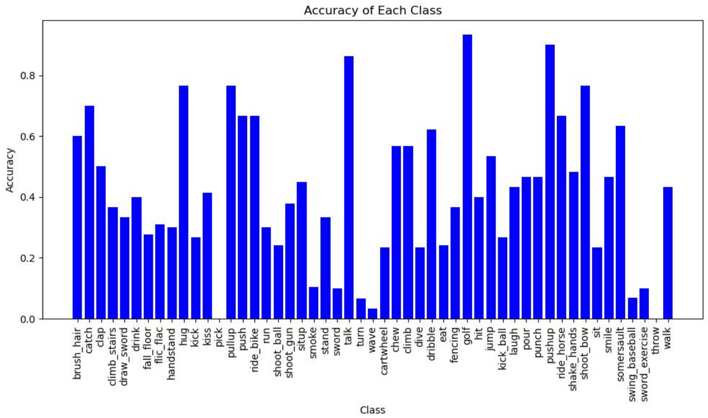


3. **Week 7:** Multimodal - Early fusion

Quantitative results:

- Improved performance regarding baseline.
- Some classes are better represented.
- The model confuses some classes.

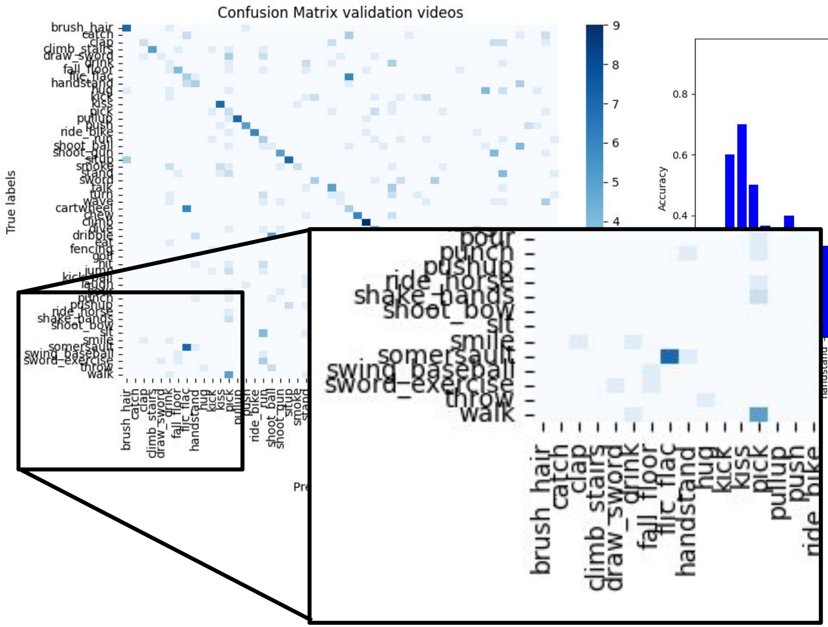


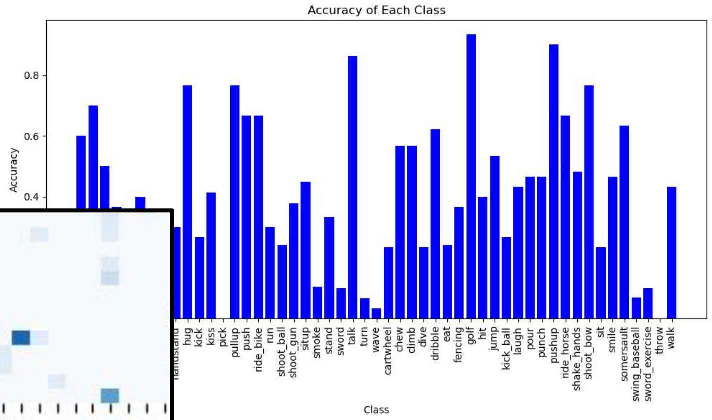


3. **Week 7:** Multimodal - Early fusion

Quantitative results:

- Improved performance regarding baseline.
- Some classes are better represented.
- The model confuses some classes.

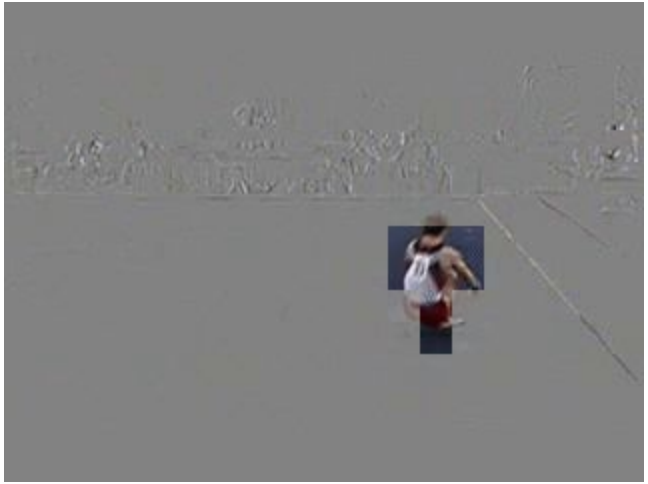




Somersault vs Flic Flac

3. Week 7: Multimodal - Early fusion

- Both are jumps!
- Kind of makes sense that the model confuses them.



Somersault

Flic Flac

3. Week 7: Multimodal - Late fusion

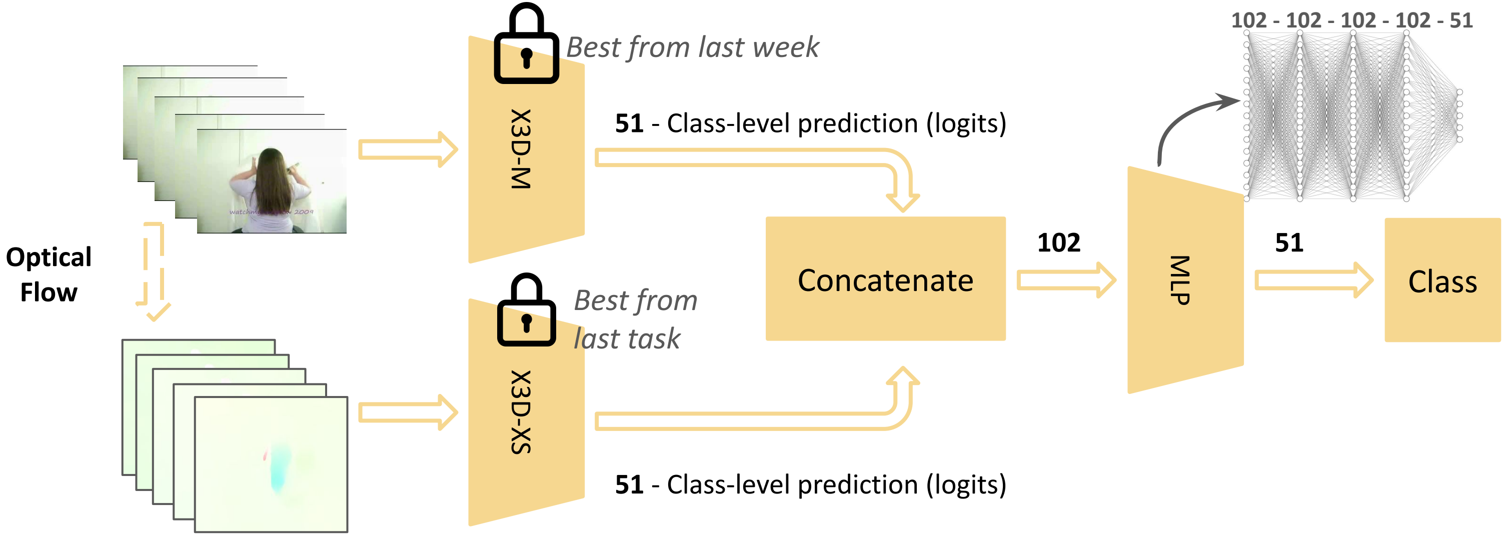
Late Fusion pipeline

1. Concatenate the **logits** of the two previously trained models for each modality.
2. MLP to fusion the predictions.
3. Predict the final class.

Motivation

- Simple approach.
- Way not to hardcode weights to aggregate the two modalities: **w1 * modality1 + w2 * modality2**.

W1 and W2 are 51 dim vectors



3. Week 7: Multimodal - Late fusion

Hyperparameter search:

- **Epochs:** 20, **30**, 40, 50.
- **Optimizer: Adam**, SGD.
- **LR:** 1e-5, 5e-5, **1e-4**, 5e-4, 1e-3.
- **Batch size:** 4, 8,**16**.

Weights & Biases

- **Overfit** on the train dataset.
- Not a good decreasing of the test loss.
- **Bad performance** in general.
- Slow to train → 9h 9m 29s.

Train acc: 0.3589

Val acc: 0.1779

Baseline

Train acc: 0.8636

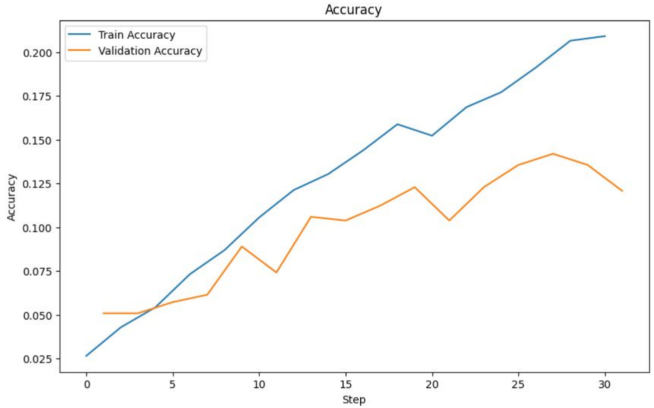
Validation acc: 0.5543

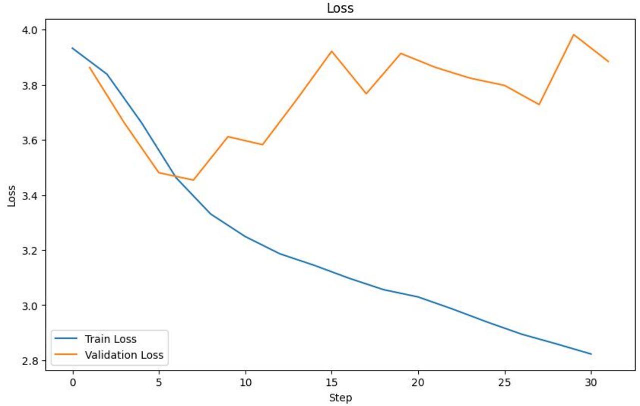
Early Fusion

Train acc: 0.2092

Validation acc: 0.1208

Late Fusion

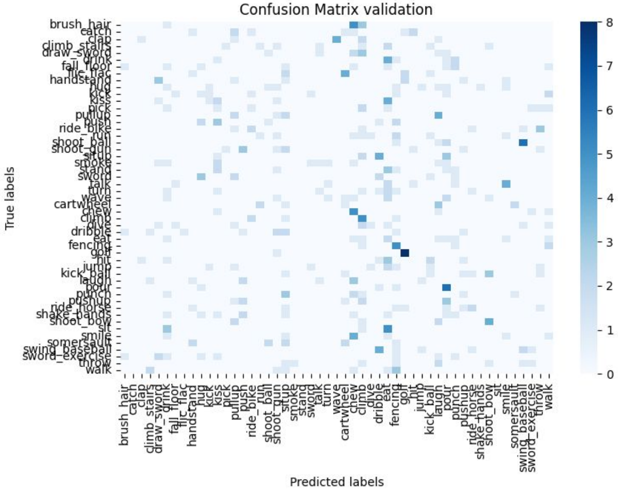


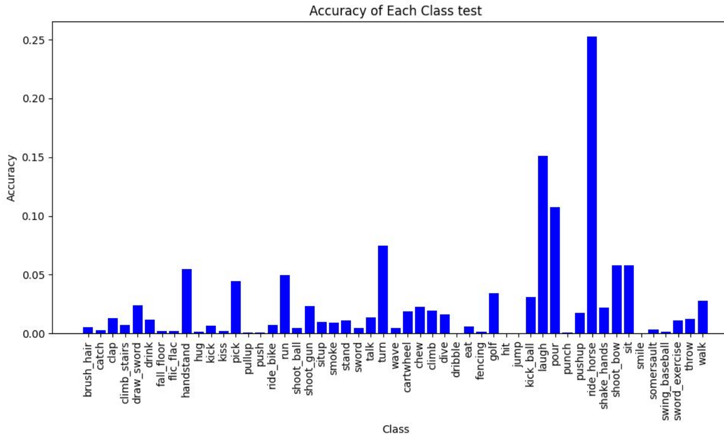


3. **Week 7:** Multimodal - Late fusion

Quantitative results:

- Poor performance, worse than baseline.
- General confusion, not biased towards one class.

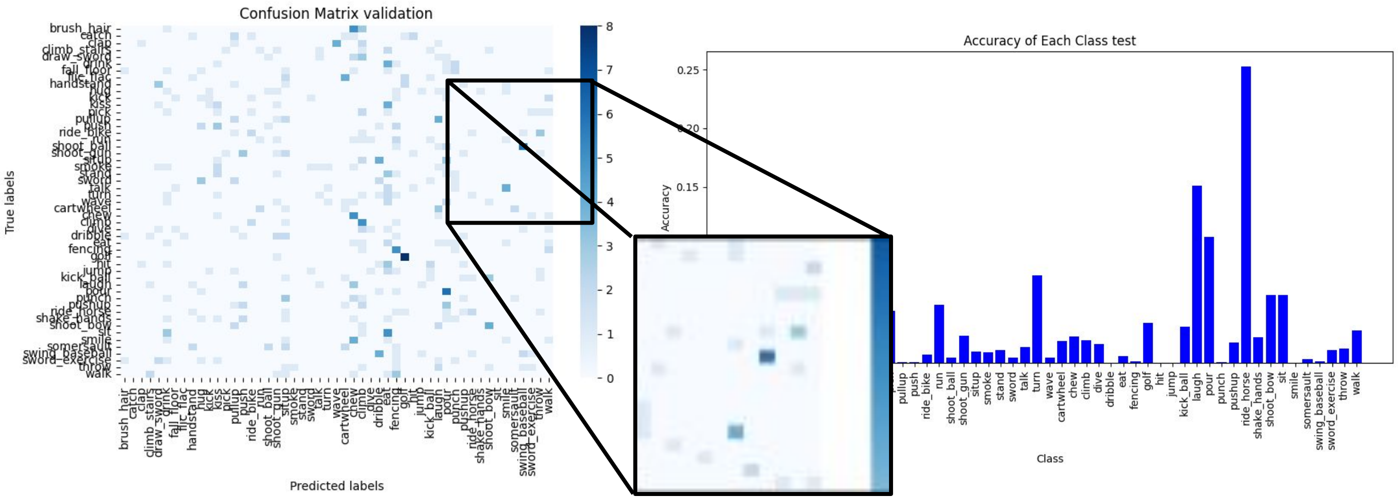




3. Week 7: Multimodal - Late fusion

Quantitative results:

- Poor performance, worse than baseline.
- General confusion, not biased towards one class.



Shoot ball vs swing baseball

3. Week 7: Multimodal - Early fusion

- Both are related to balls and sports!
- Kind of makes sense that the model confuses them.

Shoot ball

Swing baseball

3. Week 7: Multimodal - Comparison

Early Fusion - Multimodal

- ✗ Two different models.
- ✓ Can be used with pretrained weights.
- ✗ A day to train the model.
- ✗ The most data hungry model.
- ✗ No improvement from baseline.
- ✗ Didn’t learn relations from data.

Late Fusion - Multimodal

- ✗ Two different models.
- ✓ Can be used with pretrained weights.
- ✗ A long time to train.
- ✗ Very data hungry model.
- ✓ Improvement from baseline.
- ✓ Learnt relations from data.

Optical Flow

- ✓ One model only.
- ✗ Need to train from scratch or from a checkpoint.
- ✓ 35 ms to compute OF/ frame
- ✗ Data hungry.
- ✓ Improvement from baseline.
- ✓ Learnt relations from data.

4. Conclusions

- A lot of **computational power** needed to train all the models.
- **Long time** needed to train all the models.
- A lot of **memory** needed to save all the information to train the models.
- Not always adding a **new modality helps** to improve the model performance.
- Action classification is a **very hard task** as we can misclassify some of the tasks *(jumping and flic flac).*
- **Overfitting** has been present in some of the lasts experiments.

Bing Image Generator

THANK YOU!