Transformer-based Action Spotting in soccer videos

Fundamental Principles of Data Science
Master's Thesis

Artur Xarles Esparraguera

Supervisors:

- Dr. Sergio Escalera
- Dr. Albert Clapés

INTRODUCTION

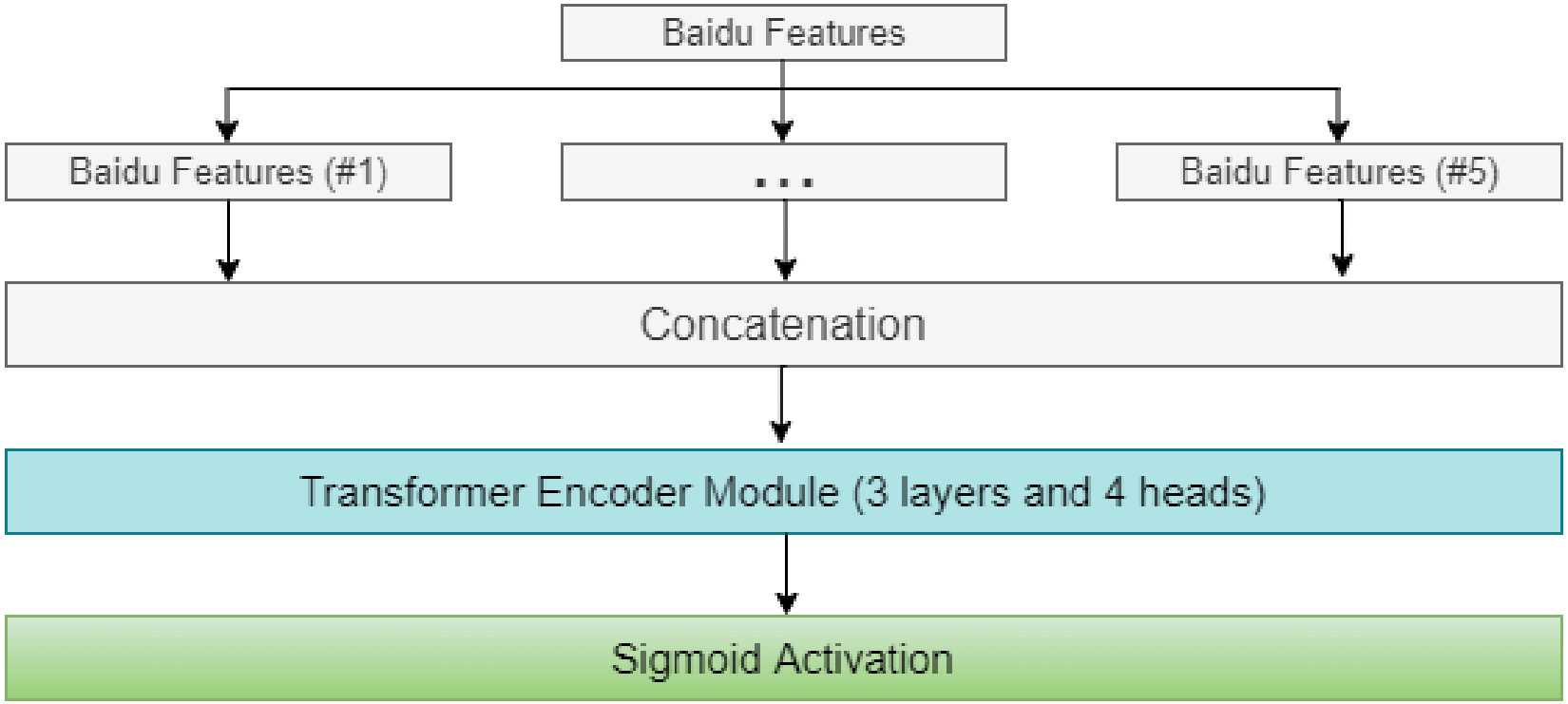
- Deep Learning in Computer Vision
- Action Spotting
- SoccerNet
- HMTAS
- 60.56% Average-mAP
- 6th position challenge (20 participants)

RELATED WORK (I)

- Small sequences of video
- Visual features from ResNet backbone (2fps)
- VLAD descriptors, CALF model [1], multi-tower CNNs
- Baseline [2]:
  - More visual features from SOTA backbones
  - Transformer-based model

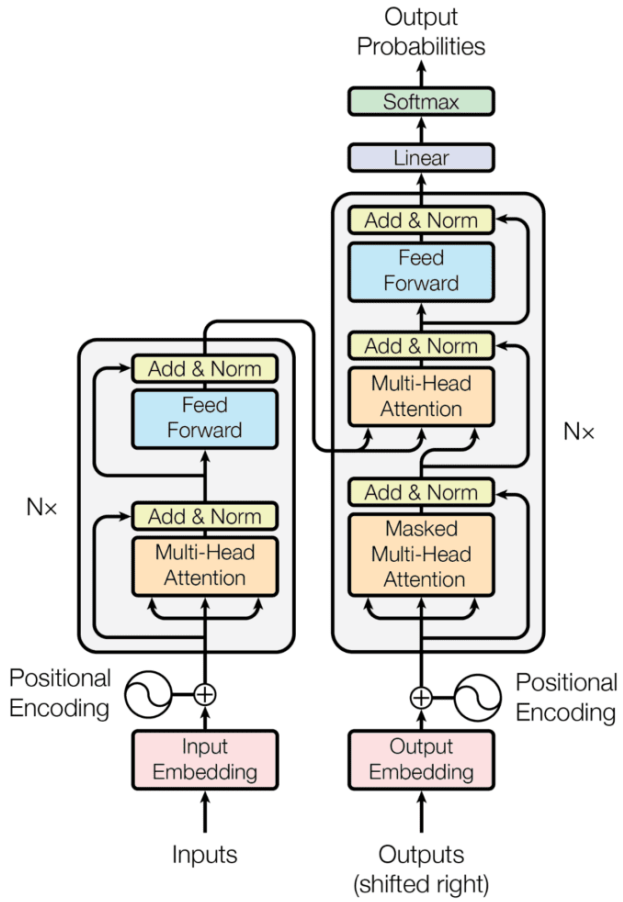
[1] A. Cioppa et al. “A context-aware loss function for action spotting in soccer videos.” In: CVPR. 2020, pp. 13126–13136.
[2] Xin Zhou et al. “Feature Combination Meets Attention: Baidu Soccer Embeddings and Transformer based Temporal Detection”. In: arXiv (2021).

RELATED WORK (II)



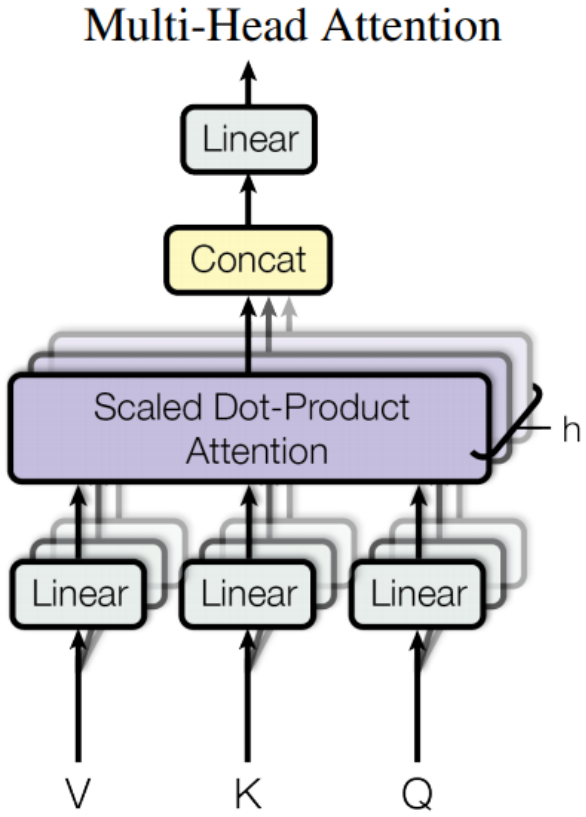
BACKGROUND – Transformer

- Neural networks
- Encoder-decoder structure
- Encoder sub-layers:
  - Self-attention mechanism
  - PFFN
- We omit decoder

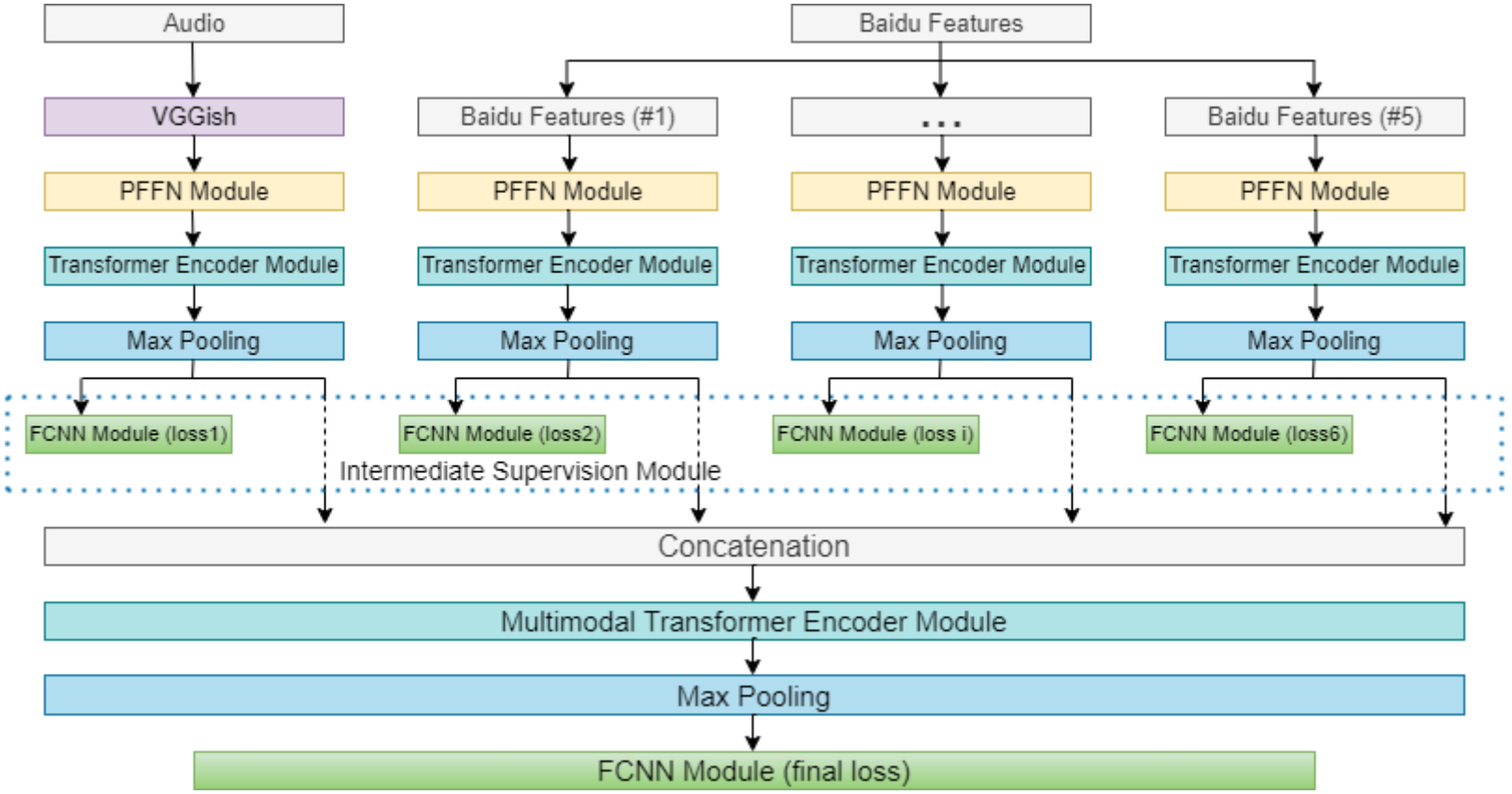


BACKGROUND – Attention

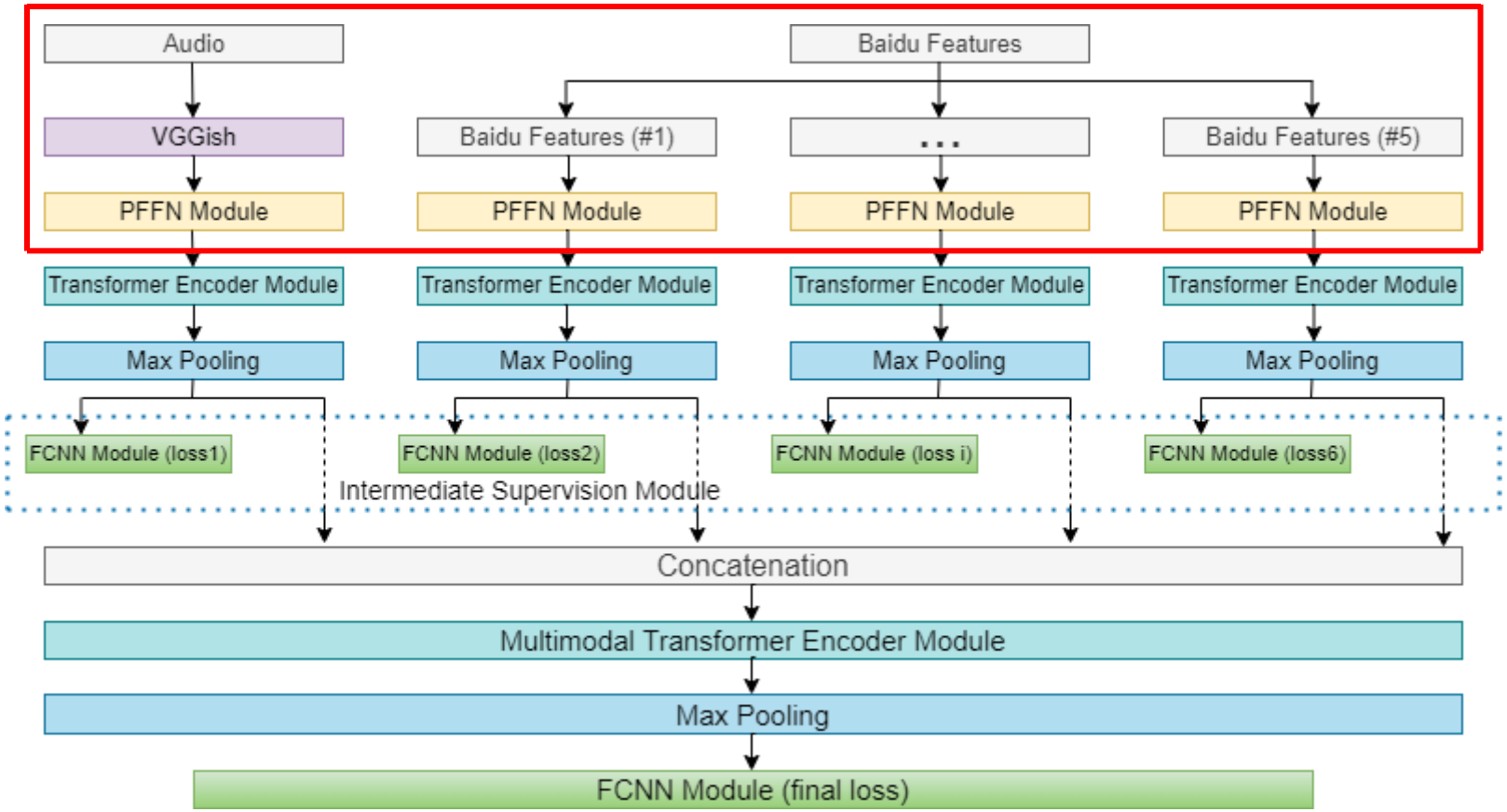
- Mimics cognitive attention
- Attention function:
  - Queries (Q)
  - Keys (K)
  - Values (V)
- Scaled Dot-Product Attention
- h attention mechanisms (projections of Q, K and V)



METHODS - HMTAS



METHODS – Embeddings (I)



METHODS – Embeddings (II)

Architecture	Type	Dimension	Pretrain
TPN	Visual	2048	K400
GTA	Visual	2048	K400
VTN	Visual	384	K400
irCSN	Visual	2048	IG65M+K400
I3D-Slow	Visual	2048	OmniSource
VGGish	Audio	512	AudioSet

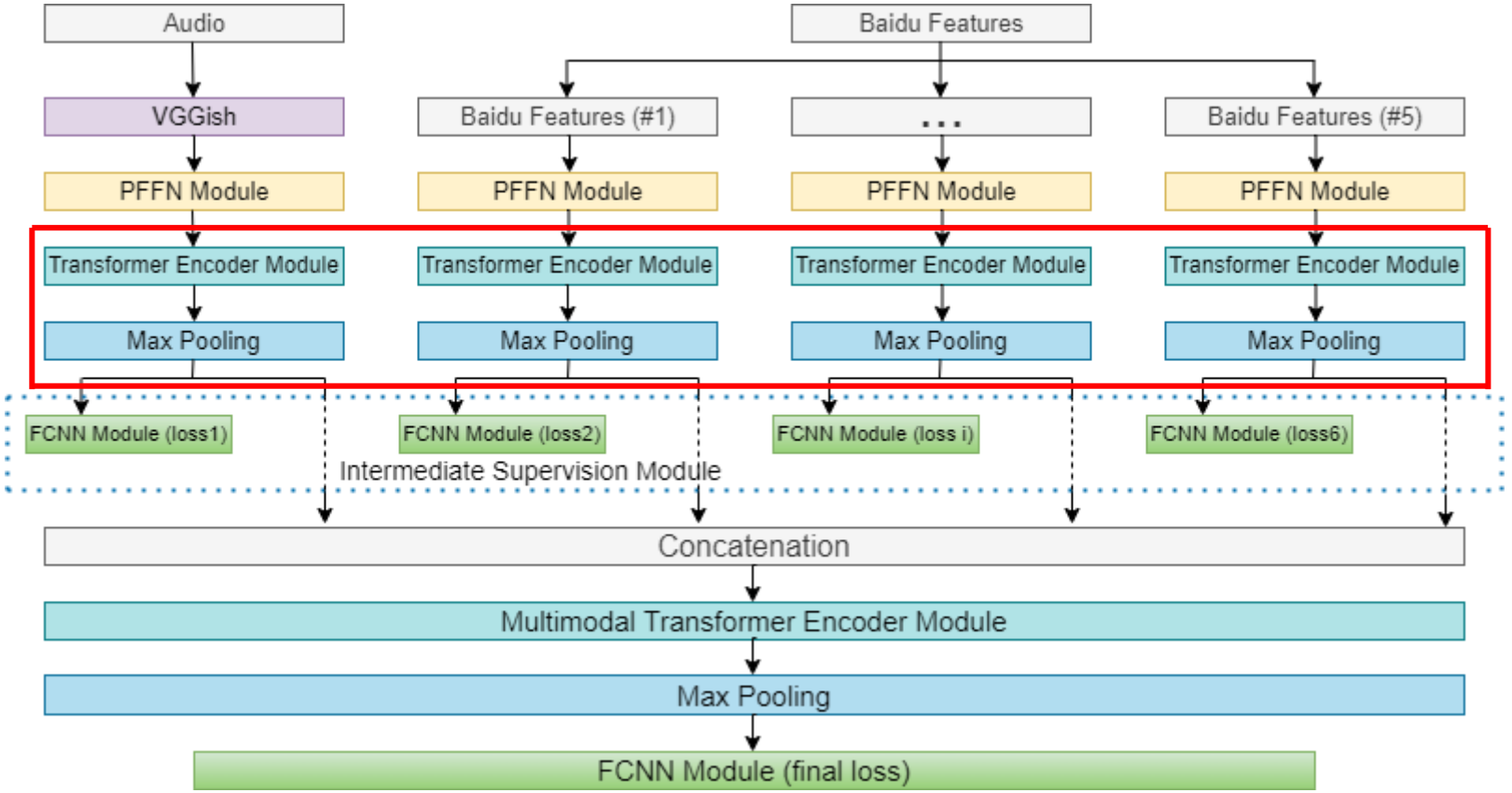
Fine-tunned on SoccerNet

Receptive field: 5 / 0.96 seconds

METHODS – Embeddings (III)

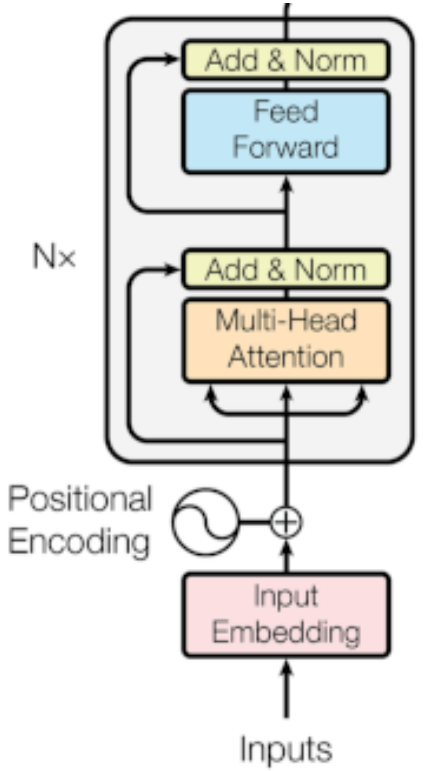
- Independent streams
- PFFN Module:
  - Point-wise Feed-Forward Network $\mathbb{R}^{L\times \mathrm{d}_t} \rightarrow \mathbb{R}^{L\times 512}$
  - Batch Normalization
  - ReLU

METHODS – Unimodal Transformer Encoders (I)

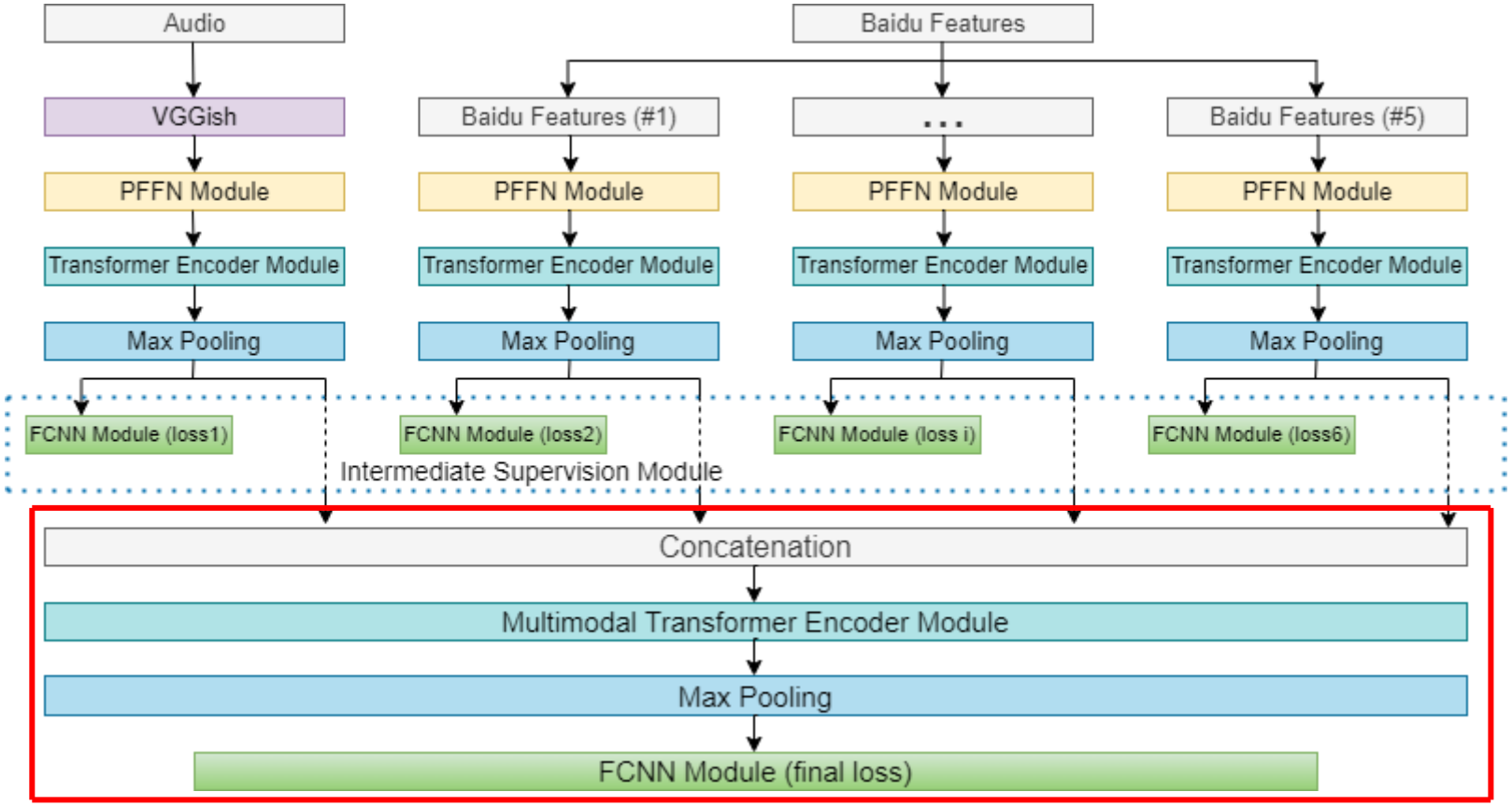


METHODS – Unimodal Transformer Encoders (II)

- 2 encoder layers with 8 heads
- Same dimensionality
- Global max-pooling $\mathbb{R}^{L\times 512} \rightarrow \mathbb{R}^{1\times 512}$

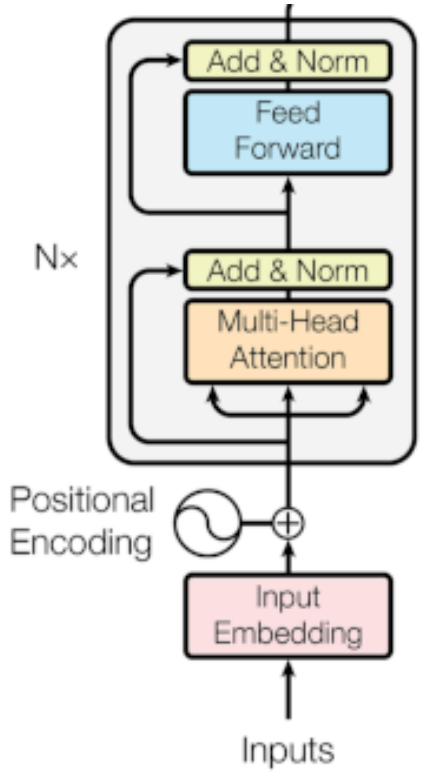


METHODS – Multimodal Transformer Encoders (I)

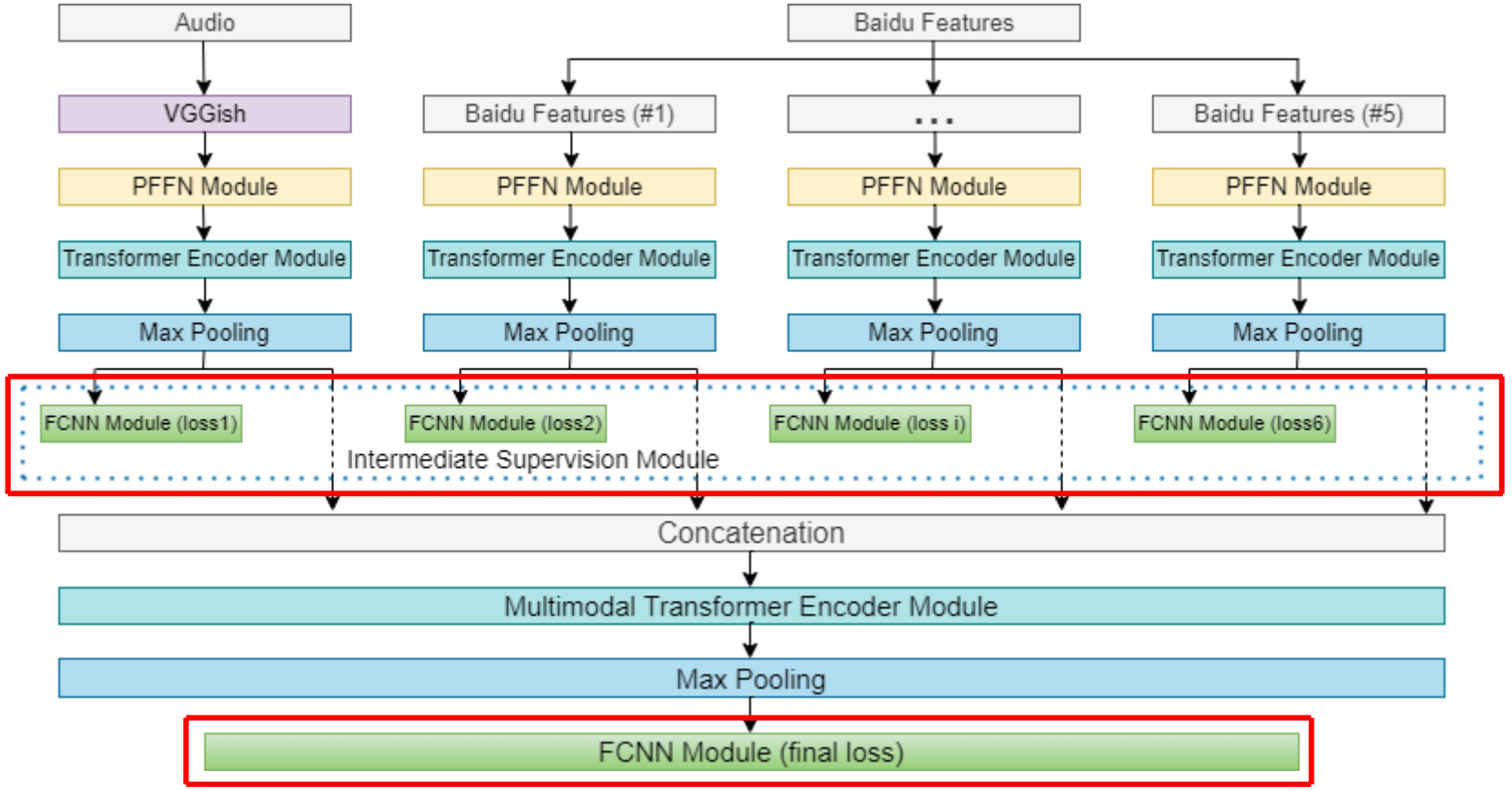


METHODS – Multimodal Transformer Encoders (II)

- Concatenation of different embeddings $\mathbb{R}^{6\times512}$
- Same configuration Transformer encoders
- Global max-pooling $\mathbb{R}^{6\times512} \rightarrow \mathbb{R}^{1\times512}$
- FCNN $\mathbb{R}^{1\times512} \rightarrow \mathbb{R}^{18}$
- Dropout and sigmoid activation



METHODS – Training Details (I)



METHODS – Training Details (II)

- Negative Log-Likelihood Loss with weights

$$NLLL(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{C}\sum_{i=1}^{C}(w_i \cdot y_i \cdot \ln \hat{y}_i + (1 - y_i) \cdot \ln(1 - \hat{y}_i))$$

- Intermediate Supervision Losses
  - 70% multimodal
  - 5% each modality

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\left(\gamma NLLL(\mathbf{y}_k, \hat{\mathbf{y}}_k^{\mathrm{MT}}) + \frac{(1-\gamma)}{E}\sum_{e=1}^{E} NLLL(\mathbf{y}_k, \hat{\mathbf{y}}_k^{e})\right)$$

METHODS – Inference

- Sliding window stride 1
- Non-Maximum Supression

FRAME	GOAL PROB.
1	0.15
2	0.20
3	0.22
4	0.20
5	0.25
6	0.30
7	0.40
8	0.50
9	0.45
10	0.55
11	0.70
12	0.65
13	0.50
14	0.30

FRAME	GOAL PROB.
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0.70
12	0
13	0
14	0

RESULTS – Dataset (I)

- 550 soccer matches
  - 300 train split
  - 100 validation split
  - 100 test split
  - 50 challenge split
- 17 different actions

Action	Penalty	Kick-off	Goal	Substitution	Offside	Shots on Target
Nº of times	173	2566	1703	2839	2098	5820
Action	Shots off target	Clearance	Ball out of play	Throw-in	Foul	Indirect FK
Nº of times	5256	7896	31810	18918	11674	10521
Action	Direct FK	Corner	Yellow Card	Red Card	YC -> RC	
Nº of times	2200	4836	2047	55	46	

RESULTS – Dataset (II)

- High intra-class variability
- Low inter-class variability
- Non visible actions

(a) Images for actions: 'Ball out of play', 'Throw-in', 'Offside' and 'Foul'.

(b) Images for actions: 'Shot on Target', 'Shot off Target', 'Direct Free-kick' and 'Indirect Free-kick'.

(c) Images for actions: 'Kick-off', 'Clearance', 'Corner' and 'Substitution'.

(d) Images for actions: 'Goal', 'Yellow card', 'Penalty' and 'Red card'.

RESULTS – Evaluation Protocols

- P1:
  - Train model → train split
  - Stop training → validation split
  - Compare models → test split

- P2:
  - Train model → train + validation + test split
  - Evaluate model → challenge split

RESULTS – Evaluation Metric

- Average-mAP: mean AP across classes and tolerances

$$AP = \sum_{t=0}^{T-1} (Recalls(t) - Recalls(t+1)) \cdot Precisions(t)$$

- Summarize precision-recall curve
- Tolerances between 1 and 5 seconds

RESULTS – Test results (I)

- M0:

  Base model

  7 second sequences

  Single stream

  Max-pooling

  FCNN

Model	*Added feature*	*Average-mAP*
Base models		
M0	FCNN ($L = 7$)	39.89%
M1	FCNN ($L = 3$)	42.59%
Transformer (L=3)		
M2	1-layer Transformer Encoder	50.02%
M3	2-layer Transformer Encoder	50.11%
Transformer model (L=3) + Incremental improvements		
M4	M3 + Hierarchy of streams	50.52%
M5	M4 + Weights in NLLL	55.05%
M6	M5 + Audio modality	56.77%
M7 (HMTAS)	M6 + Intermediate supervision	57.28%

RESULTS – Test results (II)

- M1:

3 second sequences

+2.70% Average-mAP

Less context better

Model	*Added feature*	*Average-mAP*
Base models		
M0	FCNN ($L = 7$)	39.89%
M1	FCNN ($L = 3$)	42.59%
Transformer (L=3)		
M2	1-layer Transformer Encoder	50.02%
M3	2-layer Transformer Encoder	50.11%
Transformer model (L=3) + Incremental improvements		
M4	M3 + Hierarchy of streams	50.52%
M5	M4 + Weights in NLLL	55.05%
M6	M5 + Audio modality	56.77%
M7 (HMTAS)	M6 + Intermediate supervision	**57.28%**

RESULTS – Test results (III)

- M2 & M3:

PFFN to reduce dimensionality

Transformer Encoder layers

+7.43% / +0.09% Average-mAP

Model	*Added feature*	*Average-mAP*
Base models		
M0	FCNN ($L = 7$)	39.89%
M1	FCNN ($L = 3$)	42.59%
Transformer (L=3)		
M2	1-layer Transformer Encoder	50.02%
M3	2-layer Transformer Encoder	50.11%
Transformer model (L=3) + Incremental improvements		
M4	M3 + Hierarchy of streams	50.52%
M5	M4 + Weights in NLLL	55.05%
M6	M5 + Audio modality	56.77%
M7 (HMTAS)	M6 + Intermediate supervision	57.28%

RESULTS – Test results (IV)

- M4:

Hierarchical architecture

Unimodal + Multimodal Transformer Encoder

+0.41% Average-mAP

Model	*Added feature*	*Average-mAP*
Base models		
M0	FCNN ($L = 7$)	39.89%
M1	FCNN ($L = 3$)	42.59%
Transformer (L=3)		
M2	1-layer Transformer Encoder	50.02%
M3	2-layer Transformer Encoder	50.11%
Transformer model (L=3) + Incremental improvements		
M4	M3 + Hierarchy of streams	50.52%
M5	M4 + Weights in NLLL	55.05%
M6	M5 + Audio modality	56.77%
M7 (HMTAS)	M6 + Intermediate supervision	57.28%

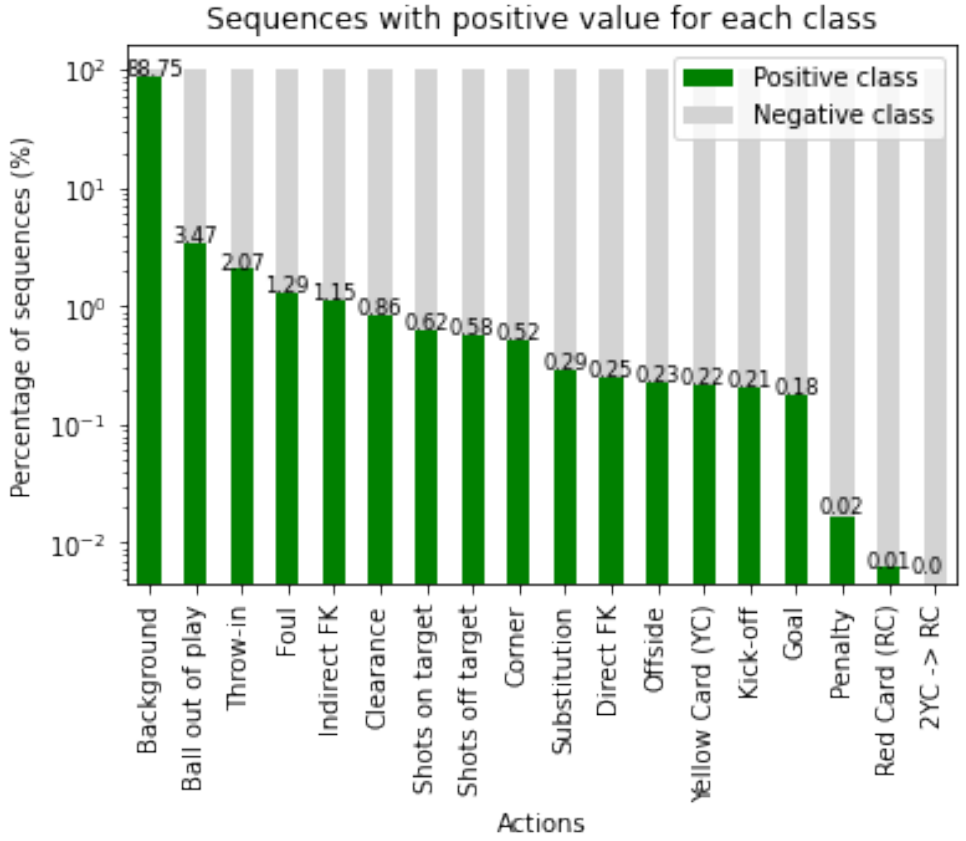
RESULTS – Test results (V)

- M5:

  NLLL with weights

  +4.53% Average-mAP

Action	*YC -> RC*	*Red Card (RC)*	*Kick-off*
mAP M4	0.00%	0.00%	49.06%
mAP M5	28.33% (+28.33%)	24.62% (+24.62%)	54.90% (+5.84%)



RESULTS – Test results (VI)

- M6:

  Audio features

  +1.72% Average-mAP

Action	*YC -> RC*	*Offside*	*Foul*	*Penalty*	*Goal*
mAP M5	28.33%	44.95%	65.80%	76.80%	73.07%
mAP M6	37.24% (+8.91%)	52.69% (+7.74%)	69.94% (+4.14%)	80.53% (+3.73%)	74.82% (+1.75%)

RESULTS – Test results HMTAS (I)

- M7:

> Intermediate Supervision Losses
>
> 57.28% (+0. 51%) Average-mAP

Action	Penalty	Kick-off	Goal	Substitution	Offside	Shots on Target
mAP	80.25%	57.26%	76.90%	47.57%	52.94%	53.33%
Action	Shots off target	Clearance	Ball out of play	Throw-in	Foul	Indirect FK
mAP	52.74%	53.95%	73.92%	69.61%	69.11%	43.66%
Action	Direct FK	Corner	Yellow Card	Red Card	YC -> RC	
mAP	58.86%	76.83%	51.33%	21.47%	34.02%	

RESULTS – Test results HMTAS (II)

- VTN best visual embeddings
- irCSN worse visual embeddings
- Audio worse than visual features

Embedding	TPN	GTA	VTN	irCSN	I3D-Slow	VGGish
Train NLLL	0.0884	0.0864	0.0527	0.1035	0.0938	0.1983
Validation NLLL	0.0926	0.0870	0.0797	0.1100	0.0976	0.1372

RESULTS – Challenge results

- Evaluation protocol P2
- 60.56% Average-mAP
- +10.98% Average-mAP respect baseline
- 6th position Soccernet 2022 Action Spotting Challenge

Model	Average-mAP
Yahoo Research	67.81%
PTS	66.73%
AS & RG	64.88%
mt_sdu_action	62.26%
Rkrystal	61.84%
HMTAS	**60.56%**
cihe	59.97%
GUC	58.71%
intro and inter	53.30%
memory	52.89%
stargazer	52.04%
Baseline	49.56%

RESULTS – Ensemble experiments

- Ensemble of different HMTAS with different sequences length (2, 3, 4, 5)

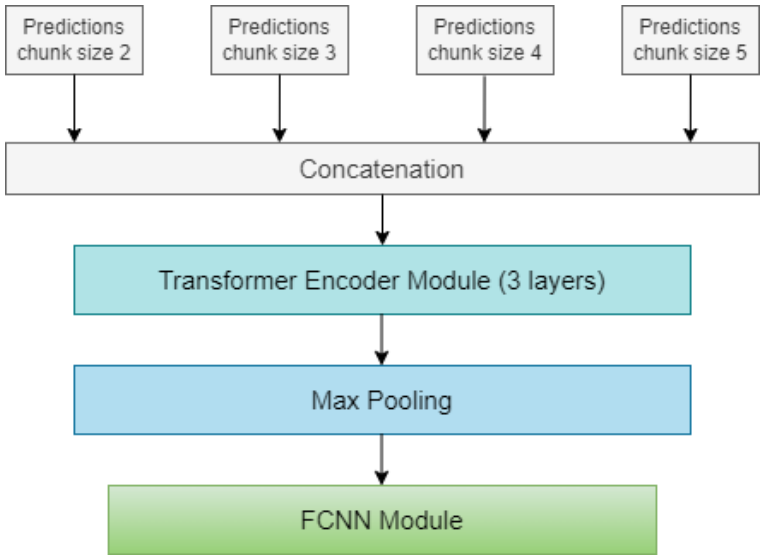
- Fusion strategies:

  Mean

  Weighted mean

  Squared weighted mean

  Learning fusion



Ensemble Strategy	Mean	Weighted mean	Squared weighted mean	Learning fusion
Average-mAP test	56.98%	56.86%	56.87%	56.92%

DISCUSSION (I)

- Action Spotting task on soccer videos
- Difficulties:
  - Unbalanced data
  - Intra-class and inter-class variability
  - Difficult actions (even for humans)
  - Non visible actions
- HMTAS:
  - 60.56% Average-mAP (+10.98% over baseline)
  - 6th position SoccerNet 2022

DISCUSSION (II)

- Main improvements:
  - ✓ Hierarchical Multimodal architecture with intermediate supervision
  - ✓ Audio features
- 2 papers under revision at ACM MMSports 2022 workshop:
  - ❖ Challenge analysis with organizers
  - ❖ HMTAS model explanation

DISCUSSION (III)

- Limitations:
  - ✕ Not end-to-end solution → Need of post-processing technique
  - ✕ Not long-term relationships → Longer sequences
- Applications:
  - ❖ Automatic information retrieval from soccer matches
  - ❖ Automatic summarization
  - ❖ Sports analytics applications

Questions?