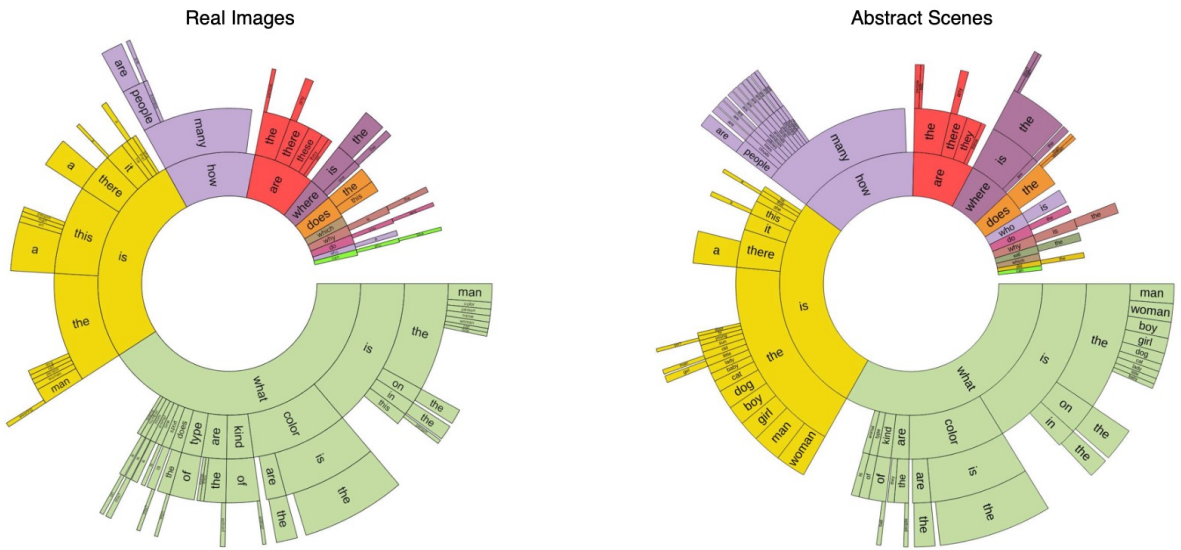
VQA [Agrawal+ (Virginia Tech), ICCV15]

- 画像と自然言語の質問を入力とし、自然言語で答えを生成するタスク
  - 250K以上の画像、760Kの質問、および約10Mの回答を含むデータセット
- real image と abstract scene が存在
- 多くのanswerは数語のみ、89.32%が1単語
  - 自動評価にも適している



Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015).
Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

VQA 2.0 [Goyal+ (Virginia Tech), CVPR17]

- 言語バイアスを大幅に削減し、バランスの取れたVQAデータセット（サイズは約2倍）
  - 「What sport is」から始まる質問のうち41%で「tennis」が正解だった
  - Do you see a ... という文に対してはyesと答えるだけで87%正解だった
- 被験者は、VQAから(Image, Question, Answer)の組 (I, Q, A)が与えられ、I に似ているが、Q に対する答えが A′ になる画像 I′ を特定
- 言語のみのモデルでは(Q, I)と(Q, I ′)を区別する根拠がない
  - 画像理解の進歩をより正確に反映
- 既存のVQAで学習したモデルはVQA 2.0ではパフォーマンスが低下

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017).
Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6904-6913).

ORT [Herdade+ (Yahoo Research), NeurIPS19]

- Image Captioningに特化したTransformer
  - 検出されたオブジェクト間の位置とサイズの関係をエンコード
- Relation box
  - TransformerにおけるΩA（右）の要素ωAに位置情報の関係を組み込む
    - (xm, ym, wm, hm)はbboxである m の中心座標、幅、高さ

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}}$$

$$\lambda(m,n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right)$$

$$\omega_G^{mn} = ReLU\left(\mathrm{Emb}(\lambda) W_G\right)$$

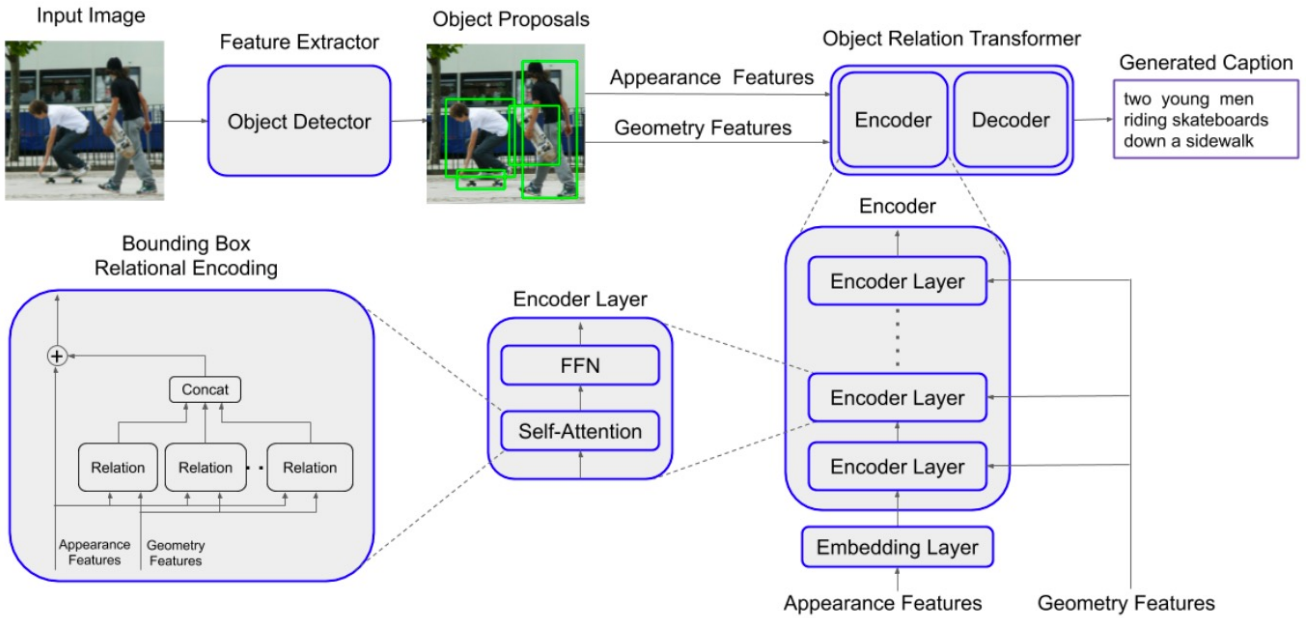
$$\omega^{mn} = \frac{\omega_G^{mn} \exp(\omega_A^{mn})}{\sum_{l=1}^{N} \omega_G^{ml} \exp(\omega_A^{ml})}$$

- Encoder Output
  - Ωはω^mnで構成されるN×N行列

$$\mathrm{head}(X) = \text{self-attention}(Q,K,V) = \Omega V$$

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W_O$$

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1) W_2 + b_2$$



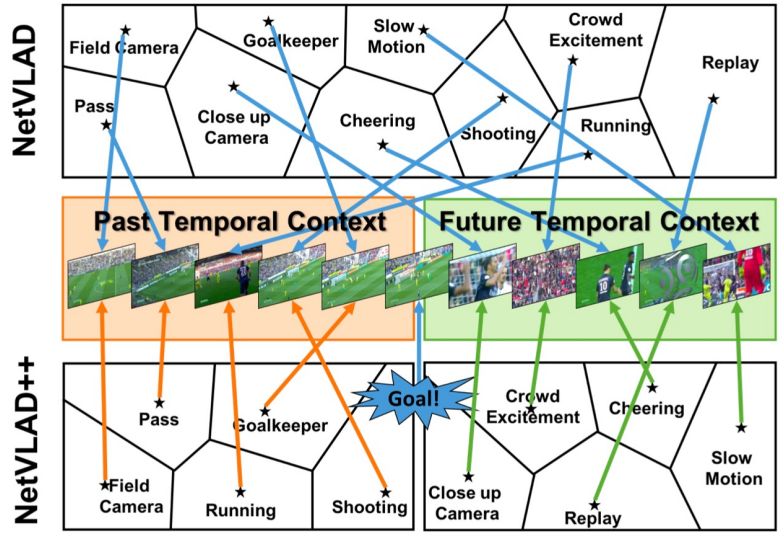
Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. arXiv preprint arXiv:1906.05963.

NetVLAD++ [Giancola+ (KAUST, Saudi Arabia), CVPR21]

- Action Spotting：タイムスタンプにおける瞬間的なイベント(=Action)を特定するタスク
- NetVLAD++：新しい特徴プーリング手法
  - 過去 [-Tb,0] と未来 [0,Ta] の特徴量を考慮
  - V=□(Vb,Va)
    - □はVbとVaのaggregation
    - Vaは[0,Ta]においてNetVLAD poolされた特徴量
- NetVLAD
  - VLADに自由度を与えたもの
  - Wk, bk, ckを最適化

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}} (\mathbf{x}_i(j) - \mathbf{c}_k(j))$$

- Average-mAP：53.4%
  - Dataset：SoccerNet-v2
  - 当時のSoTAと比較して12.7%の向上

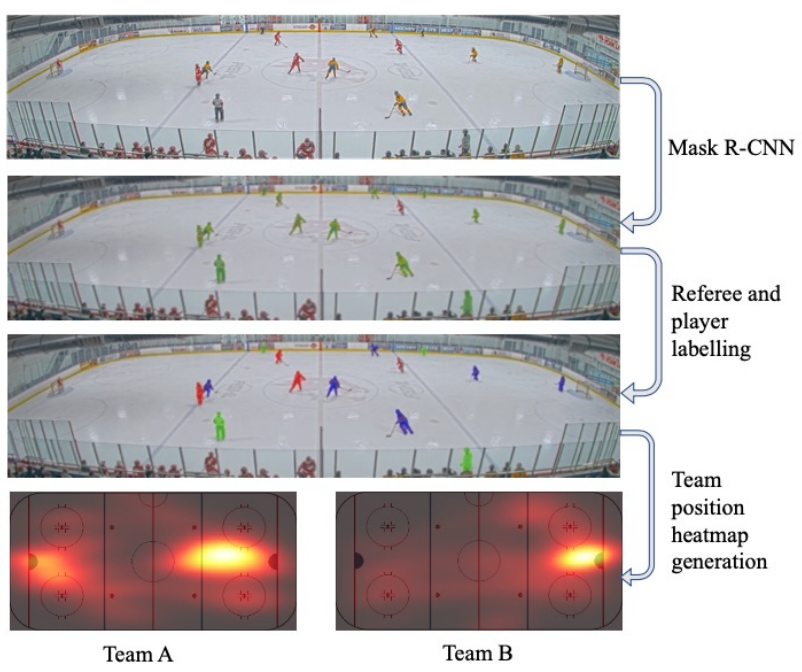


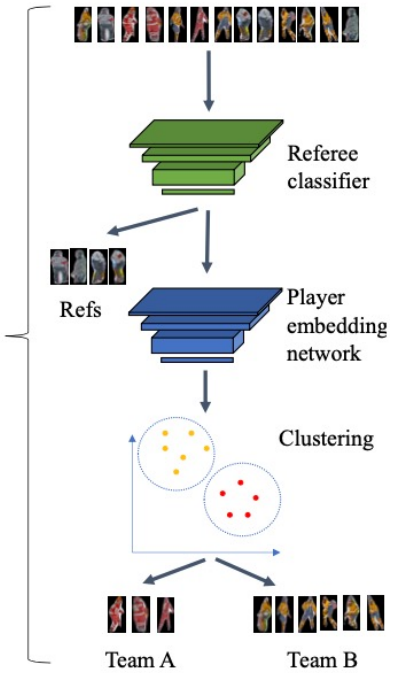
	SoccerNet-v2	visible	unshown	Ball out	Throw-in	Foul	Ind. free-kick	Clearance	Shots on tar.	Shots off tar.	Corner	Substitution	Kick-off	Yellow card	Offside	Dir. free-kick	Goal	Penalty	Yel.→Red	Red card
MaxPool [19]	18.6	21.5	15.0	38.7	34.7	26.8	17.9	14.9	14.0	13.1	26.5	40.0	30.3	11.8	2.6	13.5	24.2	6.2	0.0	0.9
NetVLAD [19]	31.4	34.3	23.3	47.4	42.4	32.0	16.7	32.7	21.3	19.7	55.1	51.7	45.7	33.2	14.6	33.6	54.9	32.3	0.0	0.0
AudioVid [41]	39.9	43.0	23.3	54.3	50.0	55.5	22.7	46.7	26.5	21.4	66.0	54.0	52.9	35.2	24.3	46.7	69.7	52.1	0.0	0.0
CALF [7]	40.7	42.1	29.0	63.9	56.4	53.0	41.5	51.6	26.6	27.3	71.8	47.3	37.2	41.7	25.7	43.5	**72.2**	30.6	0.7	0.7
NetVLAD++	**53.4**	**59.4**	**34.8**	**70.3**	**69.0**	**64.2**	**44.4**	**57.0**	**39.3**	**41.0**	**79.7**	**68.7**	**62.1**	**56.7**	**39.3**	**57.8**	71.6	**79.3**	**3.7**	**4.0**

Giancola, S., & Ghanem, B. (2021). Temporally-Aware Feature Pooling for Action Spotting in Soccer Broadcasts.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4490-4499).

Contrastive Learning for Sports Video [Koshkina+ (York Univ, Canada), CVPR21]

- チームスポーツの選手を所属チームに応じて教師なしで分類するタスク
  - ジャージの色やデザインは事前に分からないとする
  - 選手の位置に関するヒートマップを正確に計算可能
- Mask R-CNNで物体検出
- ラベル付きデータに基づいて学習した CNNで審判を分類
- 選手画像の特徴ベクトルに対し、 k-meansで2チームのcluster centerを推定
- cluster centerに基づき プレイヤーをチームに帰属
- 1フレームの教師なし学習で94%の精度
  - 500フレーム（17秒）以内に97%の精度





Koshkina, M., Pidaparthy, H., & Elder, J. H. (2021). Contrastive Learning for Sports Video: Unsupervised Player Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4528-4536).