Performers [Choromanski+ (Google), ICLR21]

- Attention行列を線形のメモリ複雑度で格納
- Fast Attention Via positive Orthogonal Random features (FAVOR+)
  - Q, Kを低ランクのQ', K'に近似
  - Q(KV)の順で計算

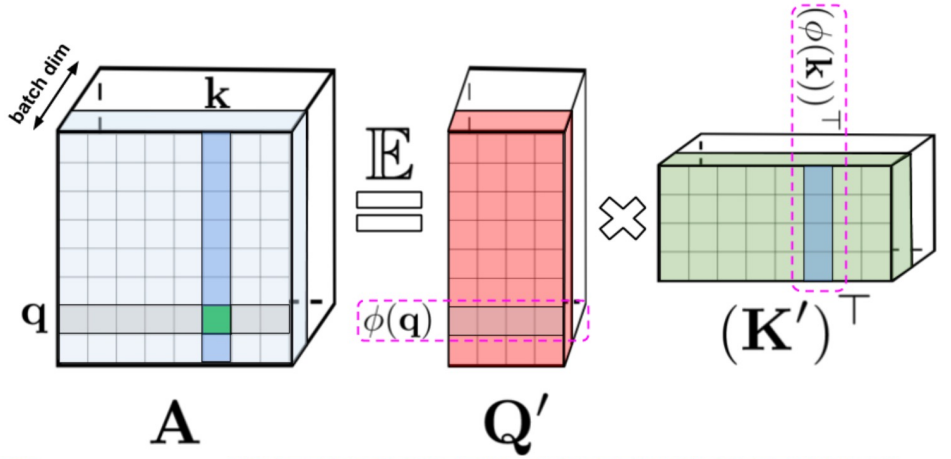
$$\mathbf{A}(i,j) = \mathrm{K}(\mathbf{q}_i^\top, \mathbf{k}_j^\top) \quad \mathrm{K}(\mathbf{x},\mathbf{y}) = \mathbb{E}[\phi(\mathbf{x})^\top \phi(\mathbf{y})]$$

$$\phi(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}}(f_1(\omega_1^\top \mathbf{x}), ..., f_1(\omega_m^\top \mathbf{x}), ..., f_l(\omega_1^\top \mathbf{x}), ..., f_l(\omega_m^\top \mathbf{x}))$$

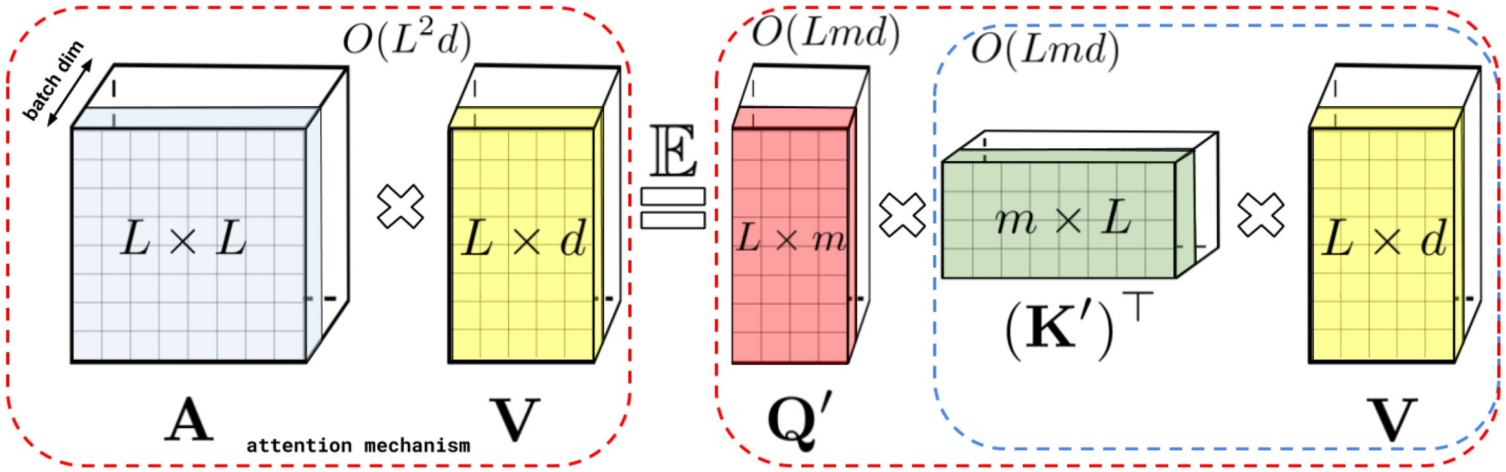
deterministic vectors ω_i or $\omega_1, ..., \omega_m \overset{\text{iid}}{\sim} \mathcal{D}$ for some distribution $\mathcal{D} \in \mathcal{P}(\mathbb{R}^d)$

$$\widehat{\mathrm{Att}}_{\leftrightarrow}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \widehat{\mathbf{D}}^{-1}(\mathbf{Q}'((\mathbf{K}')^\top \mathbf{V}))$$

$$\widehat{\mathbf{D}} = \mathrm{diag}(\mathbf{Q}'((\mathbf{K}')^\top \mathbf{1}_L))$$



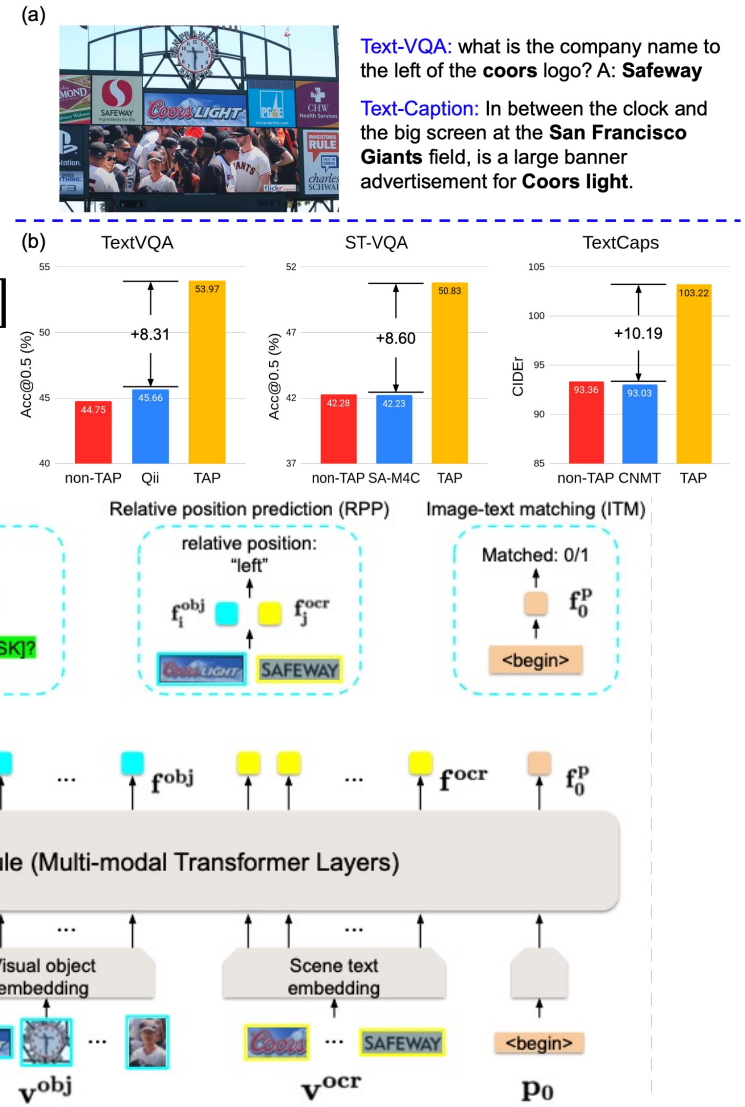
- メモリ空間
  - $O(L^2 + Ld) \rightarrow O(Lm + Ld + md)$
- 計算時間
  - $O(L^2 d) \rightarrow O(Lmd)$



Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., ... & Weller, A. (2020). Rethinking attention with performers. arXiv preprint arXiv:2009.14794.

TAP [Yang+ (Microsoft), CVPR21]

- Text-VQAやText-Captionのため、シーンテキストを事前学習
- テキストw、オブジェクト画像v^{obj}、シーンテキスト画像v^{ocr}
  開始トークンP_0
  - $w = [w^q$(質問文)$, w^{obj}$(オブジェクトのラベル)$, w^{ocr}$(シーンテキスト)$]$
    - 物体検出はFaster R-CNN、シーンテキスト検出はOCRを使用
- Fusion Module
  - 各埋め込みを連結してTransformerへ入力
- Scene-text language pre-training tasks
  - MLM：maskされたwをその位置のf^wから復元
  - ITM：f_0^Pを入力としてwが他の画像のものか判断
    - w^qのみに適応したものは殆ど効果なし
- Scene-text visual pre-training tasks
  - RPP：f^{obj}とf^{ocr}から相対的な空間的位置を予測
    - on、cover、overlap、unrelatedなど



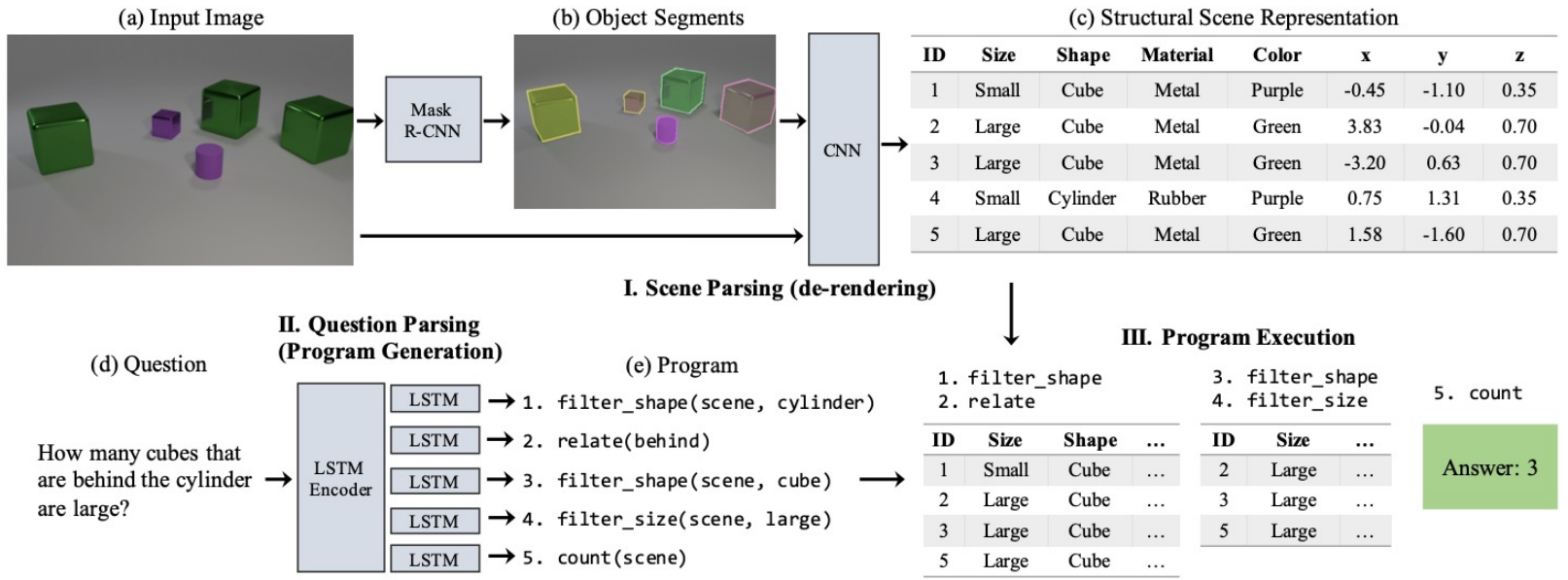
Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., ... & Luo, J. (2021). TAP: Text-Aware Pre-training for Text-VQA and Text-Caption. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8751-8761).

NS-VQA [Tenenbaum+ (MIT CSAIL), NeurIPS18]

- 画像認識や言語理解と、推論のためのsymbolic program executionを融合
  - データセット CLEVR で99.8%の精度
  - 推論時のメモリコストをSoTAから99%削減
- Scene Parser
  - Mask R-CNNで全オブジェクトのSegmentを生成
  - 色、素材、サイズ、形状などのラベルも予測
    - 元画像と対になりResNet-34に送られ、3D座標などを抽出
- Question parser
  - 質問をプログラムに変換
  - bi-LSTMのEncoder-Decoder
    - Φ_E, Φ_D : Word Embedding
- Program executor
  - プログラムを実行し答えを取得

Question parser

$$e_i = [e_i^F, e_i^B], \quad \text{where} \quad e_i^F, h_i^F = \text{LSTM}(\Phi_E(x_i), h_{i-1}^F)$$
$$e_i^B, h_i^B = \text{LSTM}(\Phi_E(x_i), h_{i+1}^B)$$
$$q_t = \text{LSTM}(\Phi_D(y_{t-1})), \ \alpha_{ti} \propto \exp(q_t^\top W_A e_i), \ c_t = \sum_i \alpha_{ti} e_i$$
$$y_t \sim \text{softmax}(W_O[q_t, c_t])$$



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

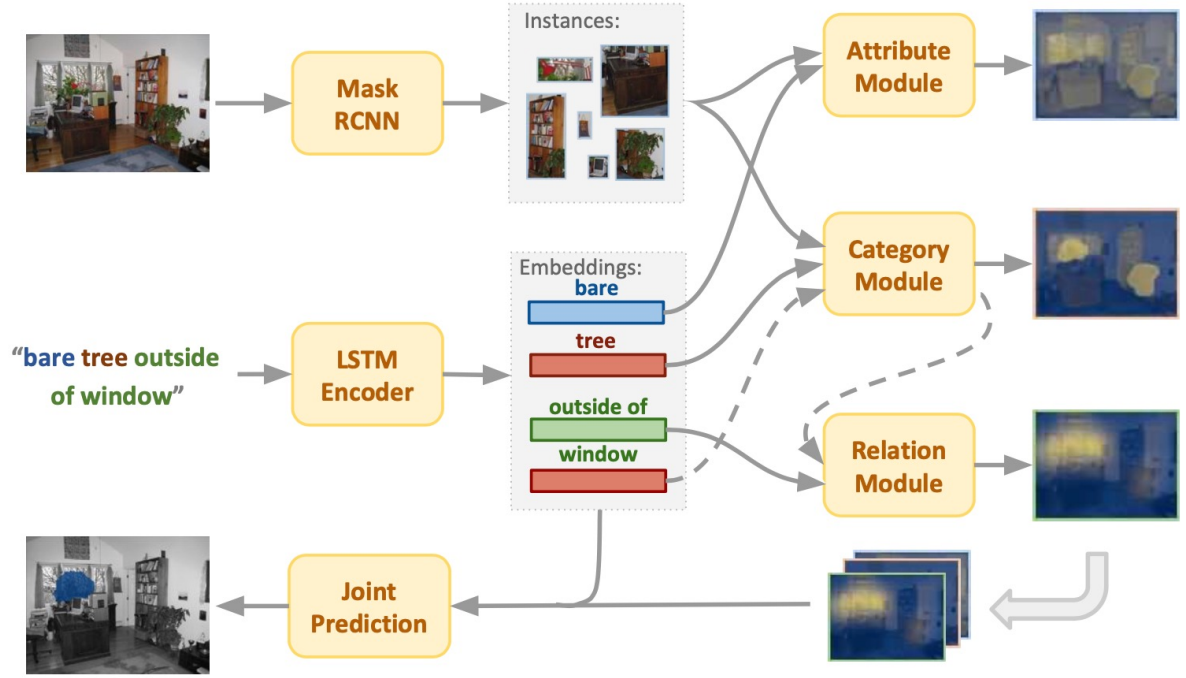
ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

ID	Size	...
2	Large	...
3	Large	...
5	Large	...

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. arXiv preprint arXiv:1810.02338.

HULANet [Wu+ (University of Massachusetts Amherst), CVPR20]

- PhraseCut Task
  - 自然言語を用いて画像領域を分離するタスク
  - データセット：VGPHRASECUTを使用
    - フレーズがCategory、Attribute、Relationを示す単語で構成
- HULANet：PhraseCutでSoTA
  - Category module：ヒートマップPを出力
    $P_{H\times W} = \sigma(a \cdot S_{H\times W} + b)$
    $A = \sigma(f(e_{cat}))\ S_{H\times W} = \sum_c A_c \cdot C_c$
    - e_{cat}はフレーズの埋め込み
    - CはMask-RCNNで各セグメントに対するスコア
  - Attribute module
    - Category moduleに分類器を追加
  - Relation Module
    - カーネルサイズ7のCNN*2を用いてscoresを取得
    - 各空間の関係に対応するフィルタを学習
  - 最終的な出力 $O = \sum_t F_t w_t$
    - 各スコアP_c, P_a, P_r、要素ごとの積$P_i \circ P_j$、バイアスによる10チャネルのF
    - 各フレーズの埋め込みを連結し、線形層と正規化を経た10次元の重みw



Wu, C., Lin, Z., Cohen, S., Bui, T., & Maji, S. (2020). Phrasecut: Language-based image segmentation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10216-10225).

JRNet [Jain+ (IIIT Hyderabad), CVPR21]

- RESの3つのベンチマークでSoTA
- Joint Reasoning Module (JRM)
  - V^p、L^p：各特徴量にPositional Encodingを加算

$$M = V^p \odot L^p$$

$$Z = \text{MultiHead}(M)$$

$$F = \text{MLP}(\text{LayerNorm}(Z))$$

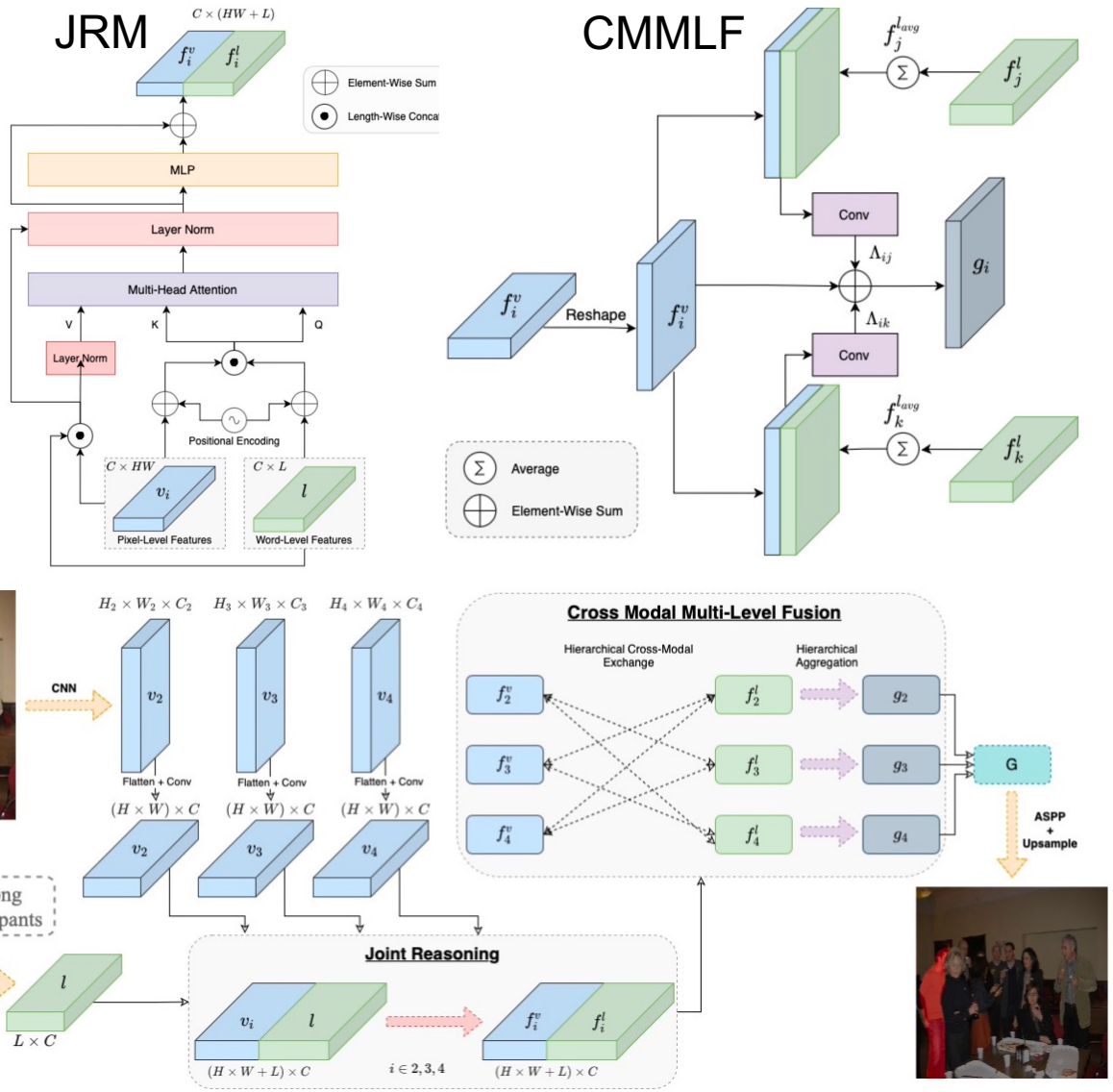
- Cross Modal Multi-Level Fusion (CMMLF)
  - ∘要素ごとの積　：チャネル方向に結合

$$\Lambda_{ij} = \sigma(Conv([f_i^v; f_j^{l_{avg}}]))$$

$$g_i = f_i^v + \sum_{j \neq i} \Lambda_{ij} \circ f_i^v$$

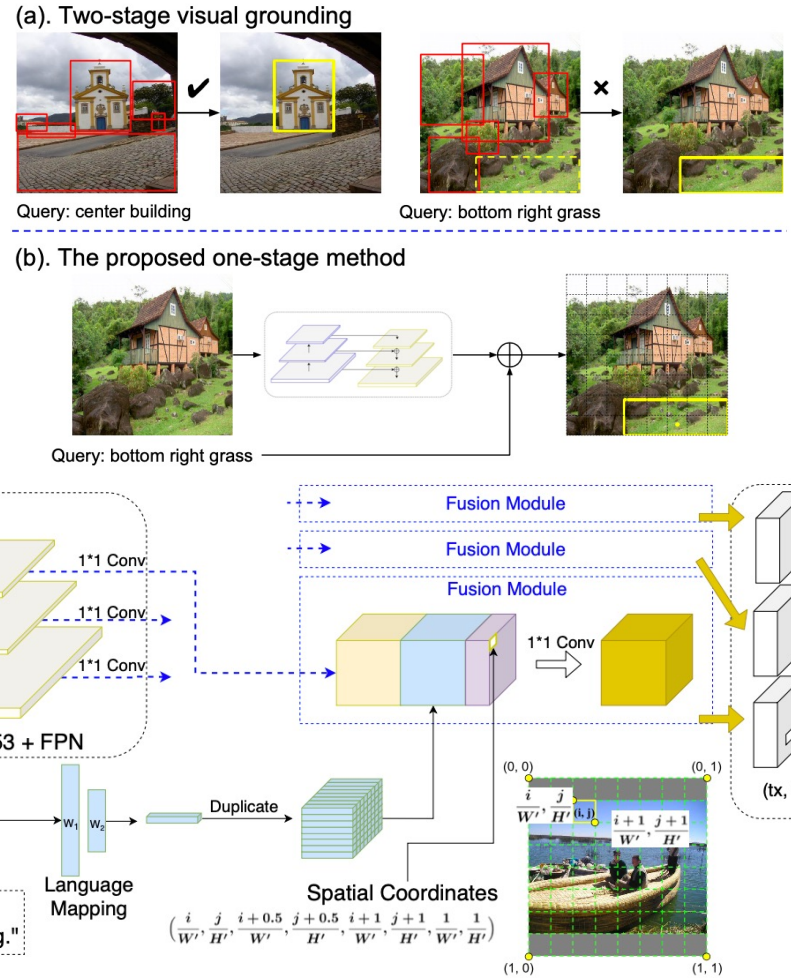
$$G = \text{Conv3D}([g_2; g_3; g_4])$$

- GはAtrous Spatial Pyramid Pooling (ASPP)デコーダとUp-samplingを経てmask Sを予測
  - Lossはbinary cross-entropy



One-Stage Approach to Visual Grounding [Yang+ (Univ. of Rochester), ICCV21]

- テキストの埋め込みをYOLOv3に融合させた1ステージモデル
  - Phrase Localization、RECにおいて精度と推論速度を両立
    - 2ステージ（bboxの候補を予測した後、テキストとの類似性により選択）のモデルの10倍の推論速度
- 画像特徴抽出： Darknet-53、Feature Pyramid Network
  - $8 \times 8 \times D_1$、,$16 \times 16 \times D_2$、$32 \times 32 \times D_3$の解像度
  - 1x1convでD=512に統一
- 言語特徴抽出：BERT+全結合層 (D=512)
- Spatial feature encoding
  - $\left(\frac{i}{W'}, \frac{j}{H'}, \frac{i+0.5}{W'}, \frac{j+0.5}{H'}, \frac{i+1}{W'}, \frac{j+1}{H'}, \frac{1}{W'}, \frac{1}{H'}\right)$
  - 各i, jに対する左上、中央、右下の座標
  - W', H'は8/16/32
- 各i, jに言語特徴量をブロードキャストし画像特徴量と結合
  - 1x1convで512+512+8→512次元
- 各i, jに対し3つのアンカーボックスを用意
  - $(8 \times 8 + 16 \times 16 + 32 \times 32) \times 3 = 4032$個のボックスに対しSoftmaxを使用してConfidenceを予測
    - 損失関数は、このSoftmaxとone-hotベクトルとの間のcross entropyを使用

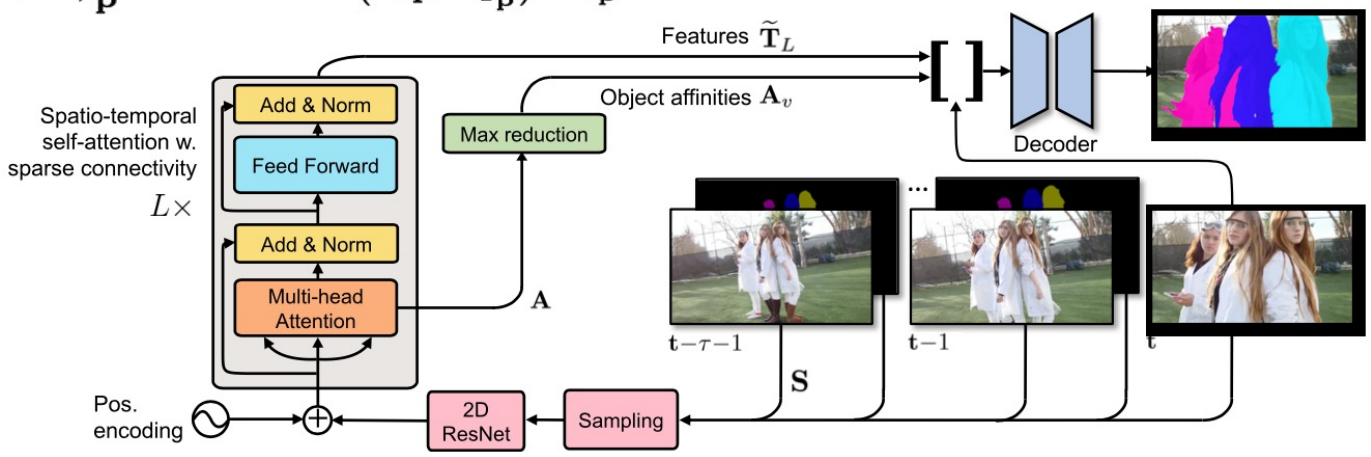


Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., & Luo, J. (2019). A fast and accurate one-stage approach to visual grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4683-4693).

SSTVOS [Duke+ (Univ. of Toronto), CVPR21]

- Transformerをベースとしたvideo Object Segmentationの手法
- 入力：長さT、サイズHxWのRGBフレーム $\mathbf{S} \in \mathbb{R}^{3\times T\times H\times W}$
  - f：ResNetで特徴抽出 $\mathbf{T} = f(\mathbf{S})\ \ \mathbf{T} \in \mathbb{R}^{C\times T\times H'\times W'}$
- τ個のフレームの埋め込みTτとPositional Encoding：Pを加算 $\widetilde{\mathbf{T}} = \mathbf{T}_\tau + \mathbf{P}$ $\mathbf{P} \in \mathbb{R}^{C\times T\times H'\times W'}$
- L層のTransformerに入力 $\widetilde{\mathbf{T}}_L = g_L \circ g_{L-1} \circ \cdots \circ g_1(\widetilde{\mathbf{T}})$
  - gはmulti-head attention
- Sparse Attentionを使用 $\mathtt{SparseAttn}(\mathbf{Q},\mathbf{K},\mathbf{V})_\mathbf{p} = \mathtt{softmax}(\mathbf{Q}_\mathbf{p}\mathbf{K}_{I_\mathbf{p}}^\intercal)\mathbf{V}_{I_\mathbf{p}}$
  - I_pは座標(i,j,k)のセット
  - どのピクセル同士がAttentionするか決定
- Attention Mapをmax pooling

$$\mathbf{A}^l \in \mathbb{R}^{|I_\mathbf{p}|\times\tau\times H'\times W'} \qquad \mathbf{A}_v^l(\mathbf{p}) = \max_{I_\mathbf{p}^o\cup\{0\}} \mathbf{A}^l$$



- Decoder
  - Segmentaion NetworkであるCFBIを使用

Duke, B., Ahmed, A., Wolf, C., Aarabi, P., & Taylor, G. W. (2021). Sstvos: Sparse spatiotemporal transformers for video object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5912-5921).