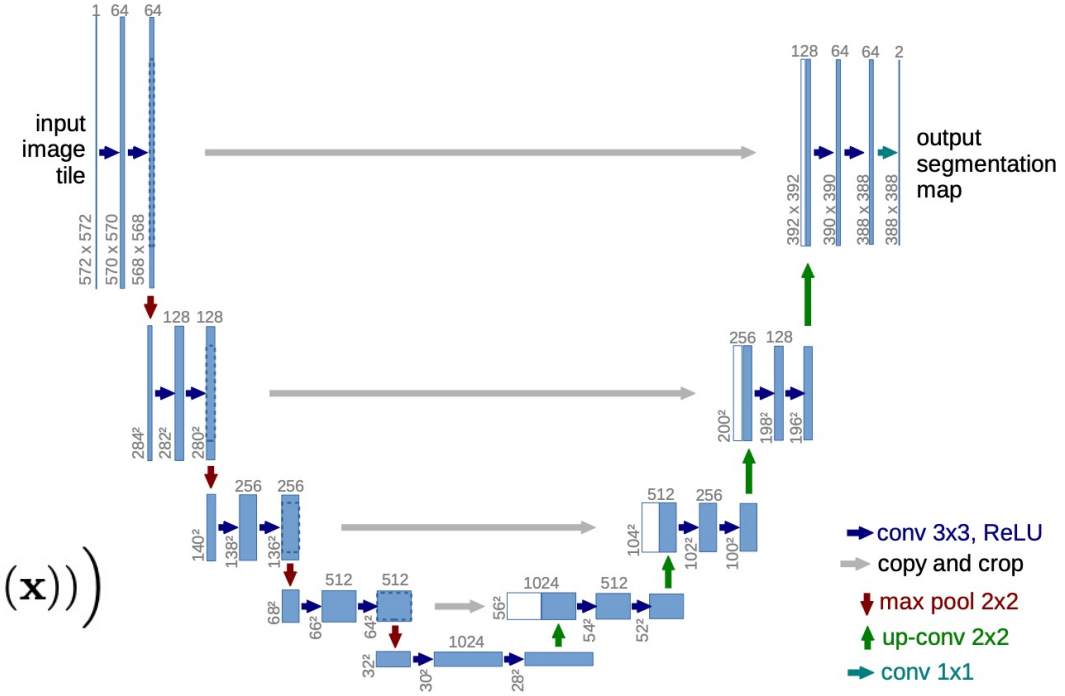
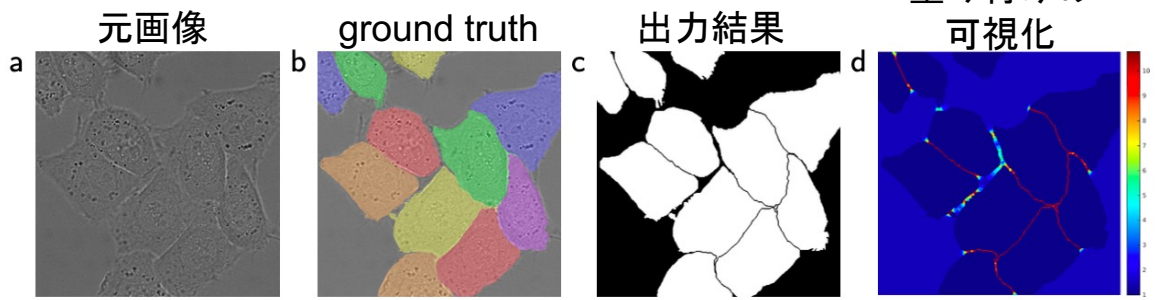
U-Net [Ronneberger+ (Univ. of Freiburg, Germany), ICMICC15]

- 少ない画像で短い学習時間での学習で正確なSegmentationを出力可能
  - Biomedical Image Segmentationで高性能
    - 30枚程の学習データから細胞のmaskを出力
    - 学習はNVidia Titan GPU（6GB）で10時間
- 左側でconvolution 右側でdeconvolution
  - マップ中央をcropし、後でconcatnate
    - 境界線のピクセルを保持
- 損失関数 $E = \sum_{\mathbf{x}\in\Omega} w(\mathbf{x}) \log(p_{\ell(\mathbf{x})}(\mathbf{x}))$

$$p_k(\mathbf{x}) = \exp(a_k(\mathbf{x})) / \left(\sum_{k'=1}^{K} \exp(a_{k'}(\mathbf{x}))\right)$$

  - x：ピクセル位置
  - $a_k(x)$：チャネルkにおけるxの活性化，k：クラス数
  - 事前に重み付け $w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) + d_2(\mathbf{x}))^2}{2\sigma^2}\right)$
    - $\sigma = 5\ pixels,\ \ w_0 = 10$
    - w_c：クラスの頻度のバランスをとるための重み
    - d_1, d_2：1番目、2番目に近い細胞の境界までの距離





Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

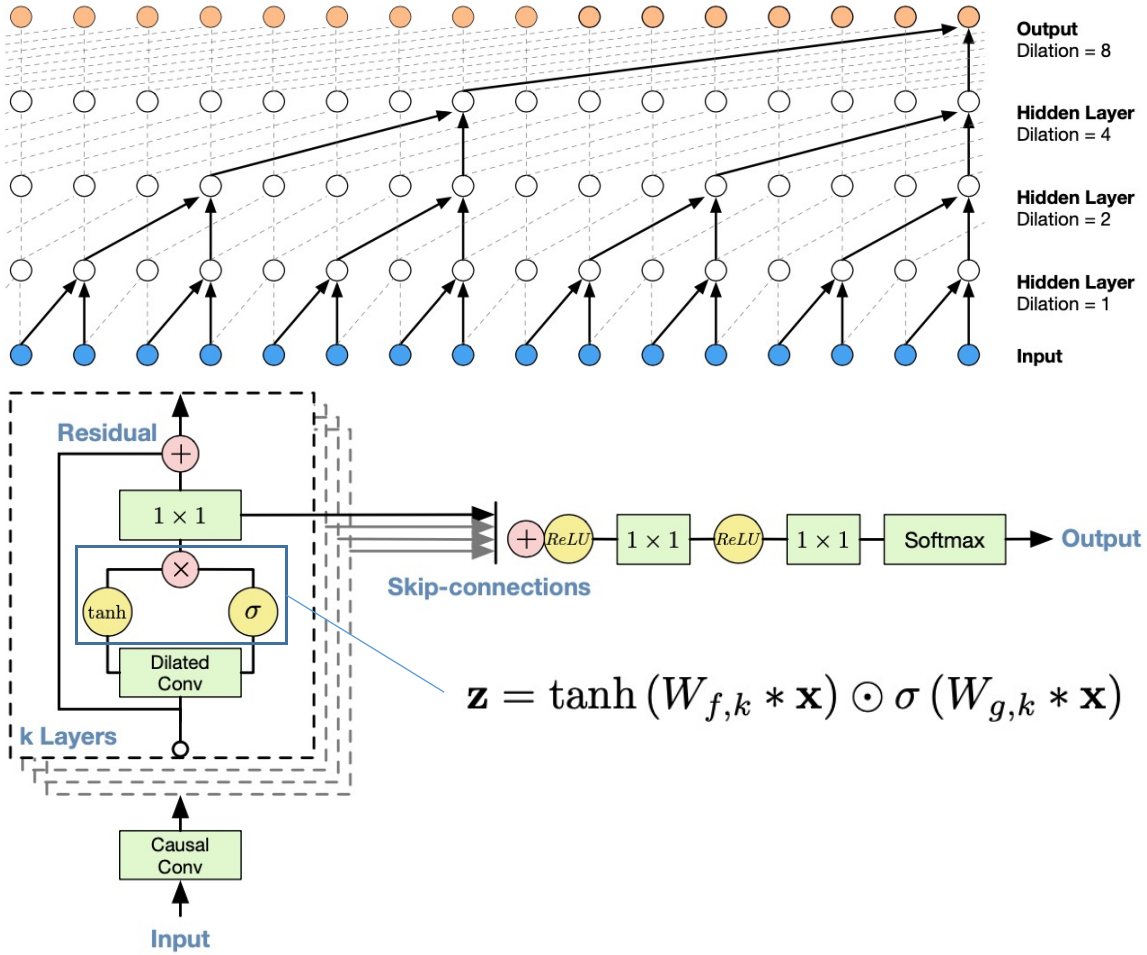
WaveNet [Oord+ (Google DeepMind), 16]

- text-to-speechで既存手法よりも自然な音声を生成
  - 時系列データを畳み込みで並列に学習することで、長い時間データ(16000Hz)を扱うことが可能
- Dilated Causal Convolution
  - dilationの幅は1,2,4, … , 512,1,2,4, … と繰り返す
- Softmax による予測
  - 16bitの音声(65536種類の値)を256個に量子化

$$f(x_t) = \mathrm{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

  - $-1 < x_t < 1, \ \mu = 255$
- Residual blockとSkip-connection
  - より深い、多様な特徴抽出
- 付加情報hを指定しxの確率分布を計算可能
  - hはテキストや話者の声など

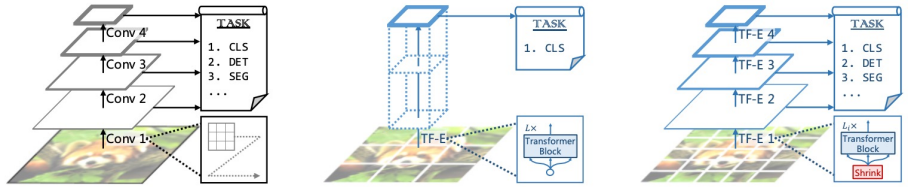
$$p(\mathbf{x}\mid\mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$



Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

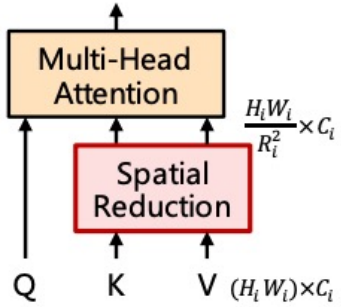
PVT [Wang+ (Nanjing Univ.), ICCV21]

- ViTにピラミッド構造を導入、任意サイズでマルチスケールの特徴マップを生成
  - 物体検出やSemantic Segmentationでも高性能
    - DETRと組み合わせも可能
  - 同パラメータ数ならCNNモデルより高性能



(a) CNNs: VGG [54], ResNet [22], *etc.* (b) Vision Transformer [13] (c) Pyramid Vision Transformer (ours)

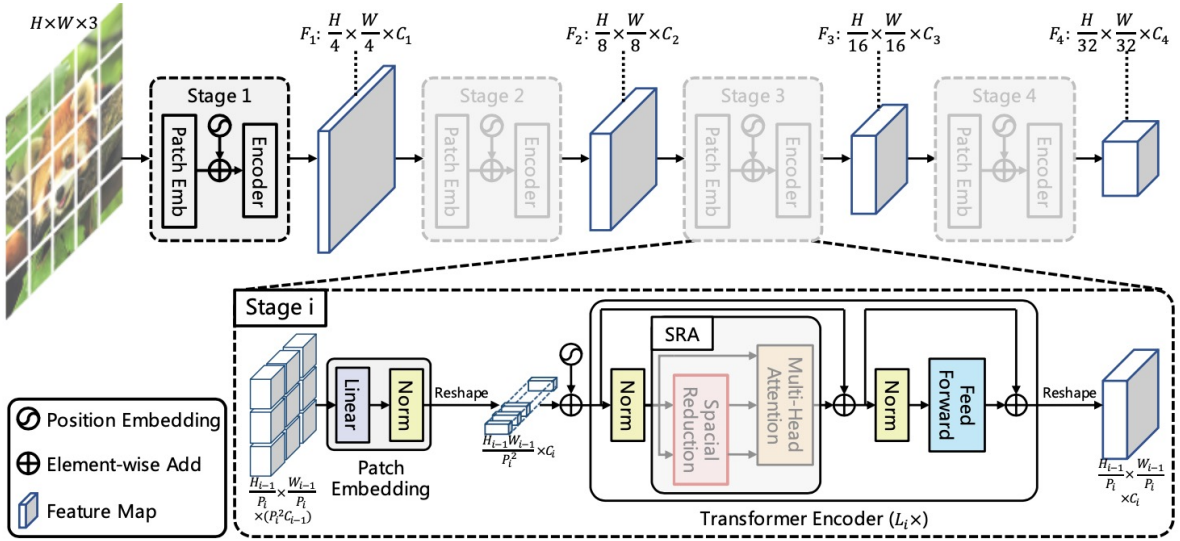
- 入力画像を分割し線形投射後、Position Embeddingを加算
- Spatial-reduction attention (SRA)
  - Multi-head attentionを改良
    - 計算量、メモリ消費を削減



$$\mathrm{SRA}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_0,\ldots,\mathrm{head}_{N_i})W^O,$$

$$\mathrm{head}_j=\mathrm{Attention}(QW_j^Q,\mathrm{SR}(K)W_j^K,\mathrm{SR}(V)W_j^V),$$

$$\mathrm{SR}(\mathbf{x})=\mathrm{Norm}(\mathrm{Reshape}(\mathbf{x},R_i)W^S)$$

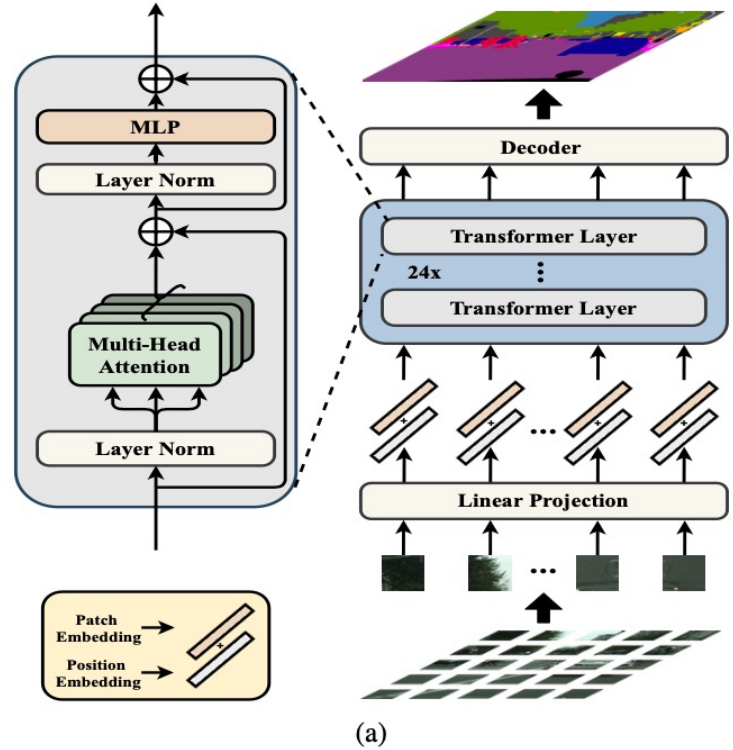


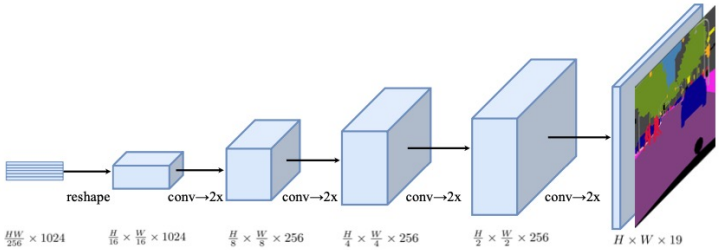
  - SRでは$x\in\mathbb{R}^{(H_iW_i)\times C_i}$を$\frac{H_iW_i}{R_i^2}\times(R_i^2C_i)$に投射後、$W_s\in\mathbb{R}^{(R_i^2C_i)\times C_i}$と乗算
    - R_iはハイパーパラメータ

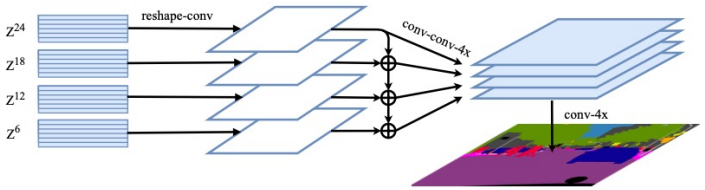
Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122.

SETR [Zheng+ (Fudan Univ.), CVPR21]

- SEgmentation TRansformer (SETR)
- EncoderにTransformerを使用
  - ADE20KデータセットでSoTA
- 入力画像を16×16のパッチに分割
  - embeddingとpositional embeddingを加算
  - TransformerでSelf-Attention
- Decoder
  - Progressive Upsampling
    - CNNでUpsamplingを繰り返す
  - Multi-Level feature Aggregation
    - M個の層に分割してデコーダに入力
    - 途中の特徴量を足し合わせながら畳み込み







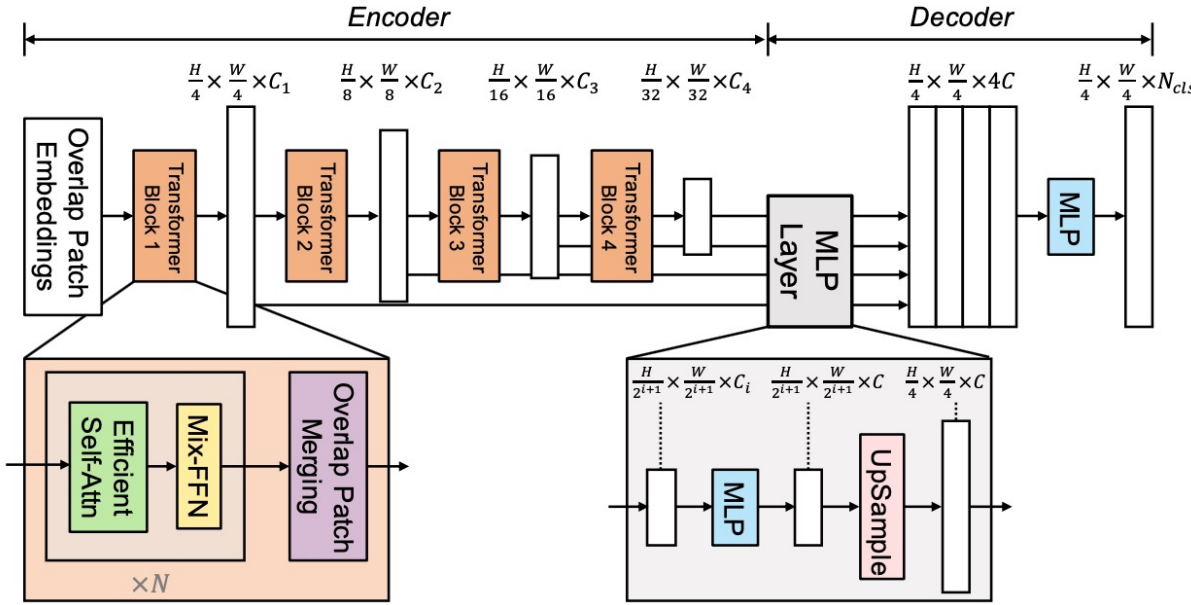
Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6881-6890).

SegFormer [Xie+ (The Univ. of Hong Kong), 21]

- Semantic SegmentationのためのTransformer
  - モデルサイズ、実行時間、精度で優位性を実証、また高いロバスト性
    - ADE20Kでは既存手法よりも4倍小さい上にmIoUでSoTA
  - positional encodingを使用しない
- 階層的なTransformer Encoder
  - $H \times W \times 3$の入力画像を$4 \times 4 \times C_1$のパッチに分割
- Efficient Self-Attention: $O(\frac{N^2}{R})$

$$\hat{K} = \text{Reshape}(\frac{N}{R}, C \cdot R)(K)$$
$$K = \text{Linear}(C \cdot R, C)(\hat{K}),$$

  - Keyを$N \times C$から$\frac{N}{R} \times (C \cdot R)$にReshape、$C \cdot R$を$C$に線形変換
    - $N = H \times W$, Rは$[64, 16, 4, 1]$
- Mix-FFN $\quad \mathbf{x}_{out} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(\mathbf{x}_{in})))) + \mathbf{x}_{in}$
  - FFNに3×3 Convを使用
- Overlap Patch Merging
  - i層のマップ$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$を$i+1$層に変換
  - K:パッチサイズ　S:隣接2パッチ間のストライド　P:パディングサイズ　を定義
    - オーバラップしてパッチを結合、パッチ周辺の連続性を保持
- 軽量なAll-MLP Decoder（右式）
  - 複数の特徴量マップを集約しマスクを生成



$$\hat{F}_i = \text{Linear}(C_i, C)(F_i), \forall i$$
$$\hat{F}_i = \text{Upsample}(\frac{W}{4} \times \frac{W}{4})(\hat{F}_i), \forall i$$
$$F = \text{Linear}(4C, C)(\text{Concat}(\hat{F}_i)), \forall i$$
$$M = \text{Linear}(C, N_{cls})(F),$$

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv preprint arXiv:2105.15203.

CMPC [Huang+ (Univ. of Chinese Academy of Sciences), CVPR20]

- 参照するEntityを強調するCMPC + 多段階の特徴間で情報交換を行うTGFE
  - Referring Image Segmentationで当時SoTA
- 単語をEntity、Attribute、Relation、Unnecessaryに分類

$$p_t = [p_t^{ent}, p_t^{attr}, p_t^{rel}, p_t^{un}]$$
$$p_t = softmax(W_2\sigma(W_1 l_t + b_1) + b_2)$$

  - l_t : t番目の単語のembedding　σ : sigmoid
- Entity Fusion　$q = \sum_{t=1}^{T}(p_t^{ent} + p_t^{attr})l_t$
  - rはハイパーパラメータ　$M_i = (qW_{3i}) \odot (XW_{4i})$　$M = \sum_{i=1}^{r} M_i$
- Cross-Modal Progressive Comprehension (CMPC)
  - 連結グラフをGraph Convolution　$R = \{r_t\}_{t=1}^{T}$　$r_t = p_t^{rel} l_t$
    - Mは頂点の特徴量Mgに線形変換　$B = (M_g W_5)(RW_6)^T$,
    - Aは隣接行列、Iは単位行列　$B_1 = softmax(B)$,
  - $s, X, \overline{M_g}$を連結しYとする　$B_2 = softmax(B^T)$,

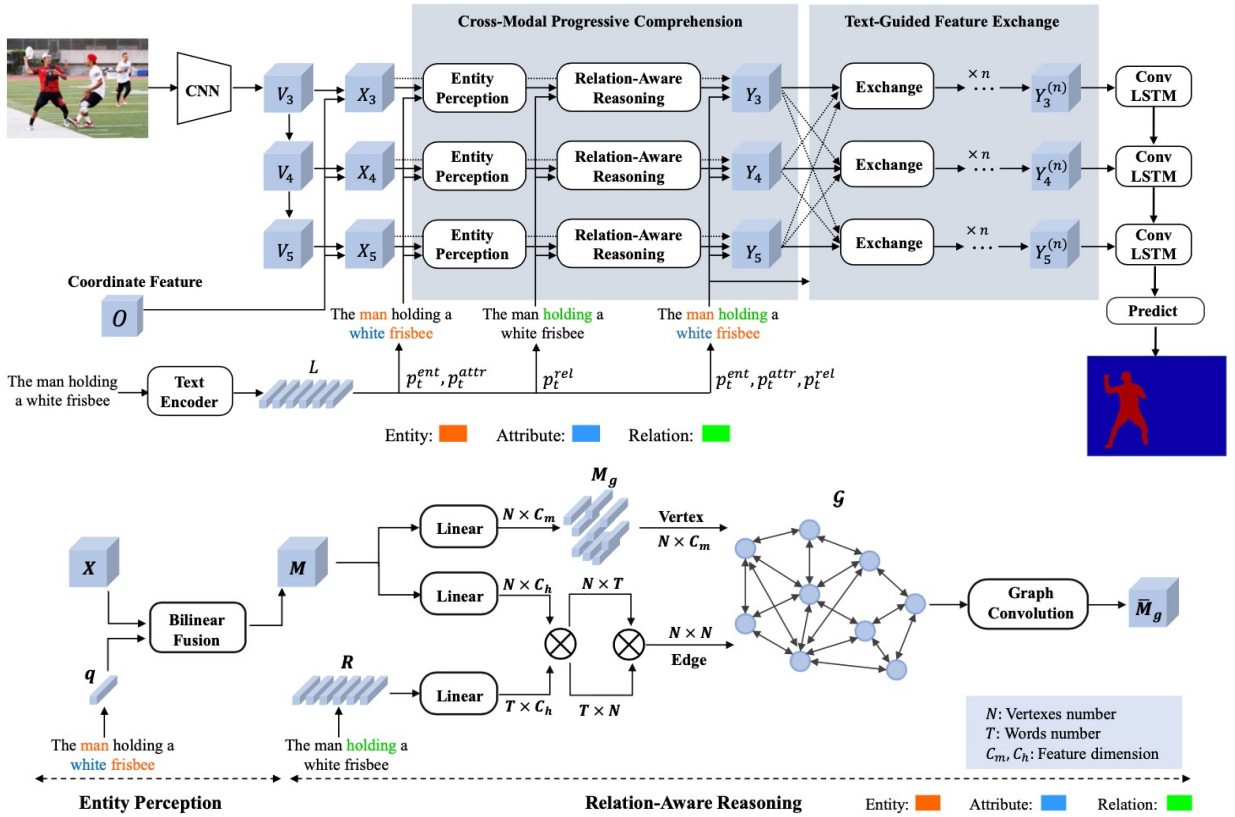
$$s = \sum_{t=0}^{T}(p_t^{ent} + p_t^{attr} + p_t^{rel})l_t$$

$$A = B_1 B_2,$$
$$\bar{M}_g = (A + I)M_g W_7$$

- Text-Guided Feature Exchange (TGFE)
  - nラウンドの特徴交換を行う、c_iはg_i, sを連結
  - $Y_3^{(n)}, Y_4^{(n)}, Y_5^{(n)}$ をConvLSTMで融合し、最終出力

$$Y_i^{(k)} = \begin{cases} Y_i^{(k-1)} + \sum_{j \in \{3,4,5\} \setminus \{i\}} \sigma(c_i^{(k-1)}) \odot Y_j^{(k-1)}, k \geq 1 \\ Y_i, k = 0 \end{cases}$$

$$g_i^{(k-1)} = \Lambda_i^{(k-1)} Y_i^{(k-1)},$$
$$\Lambda_i^{(k-1)} = (sW_8)(Y_i^{(k-1)} W_9)^T,$$



Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., ... & Li, B. (2020). Referring image segmentation via cross-modal progressive comprehension. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10488-10497).

Locate-then-Segment [Jing+ (CRIPAC), CVPR21]

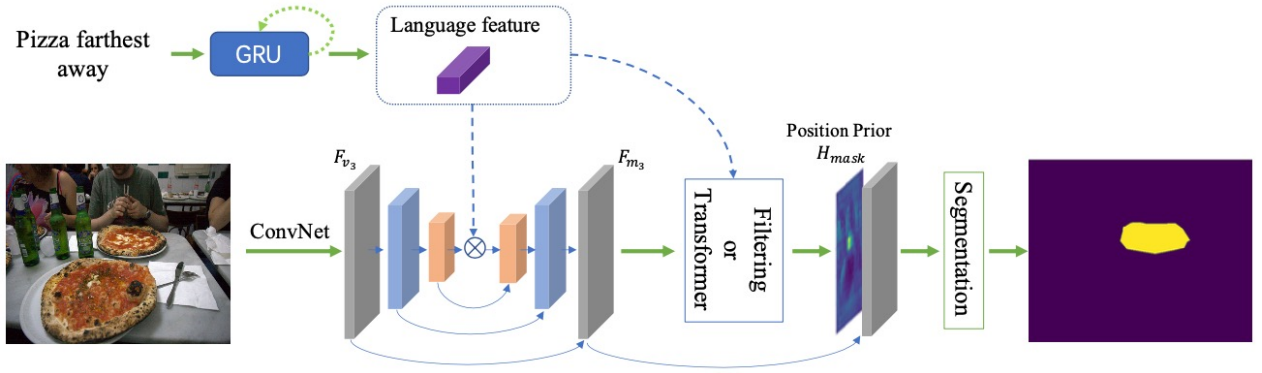
- RISを参照物体の予測とマスクの生成という2つの連続したタスクに分解
- ConvNets、GRUでそれぞれ特徴抽出
- multi-modal tensorを作成

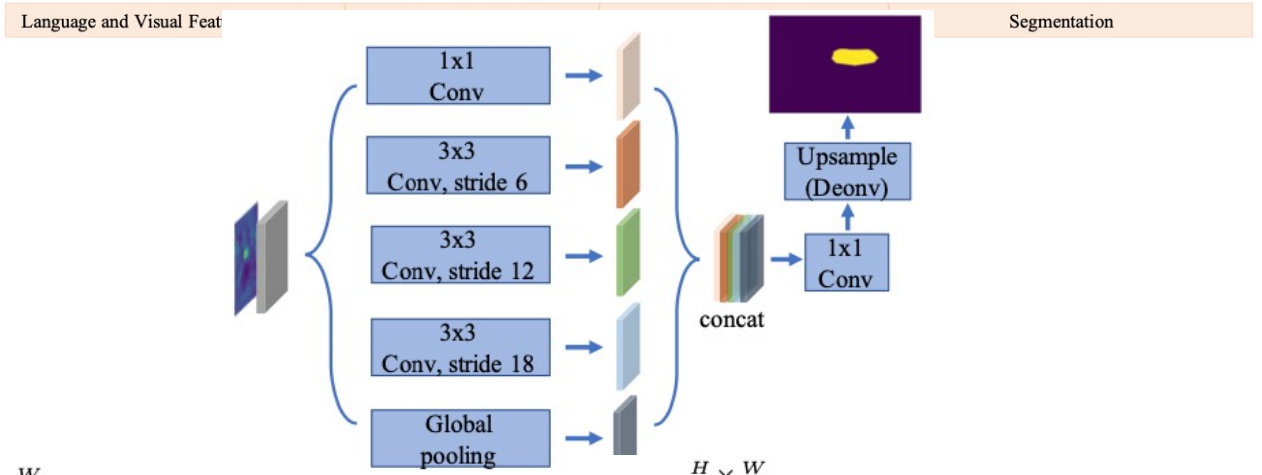
$$f^l_{m_1} = g(f^l_{v_1} W_{v_1}) \cdot g(f_{text} W_t) \quad F'_{m_{i-1}} = UpSample(F_{m_{i-1}})$$
$$F_{m_i} = concat(g(F'_{m_{i-1}} W_{m_{i-1}}), g(F_{v_i} W_{v_i}))$$

- 2種類のLocalization
  - 言語情報から生成したカーネルKでconv

    $K = f_{text} W_k \quad H_{mask} = conv(K, F_{m_3})$
  - TransformerのDecoderを使用
    - f_{text}をEncoderの出力とみなす

      $H_{mask} = decoder(F_{m_3}, f_{text})$
  - 性能はTransformerが若干良い
- ASPP Decoderでマスクを生成





$$L_{seg} = \sum_{l=1}^{\frac{H}{4} \times \frac{W}{4}} [y_l \log(p_l) + (1 - y_l) \log(1 - p_l)] \quad L_{loc} = \sum_{l=1}^{\frac{H}{8} \times \frac{W}{8}} [y_l \log(h_l) + (1 - y_l) \log(1 - h_l)]$$

Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., & Tan, T. (2021). Locate then Segment: A Strong Pipeline for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9858-9867).