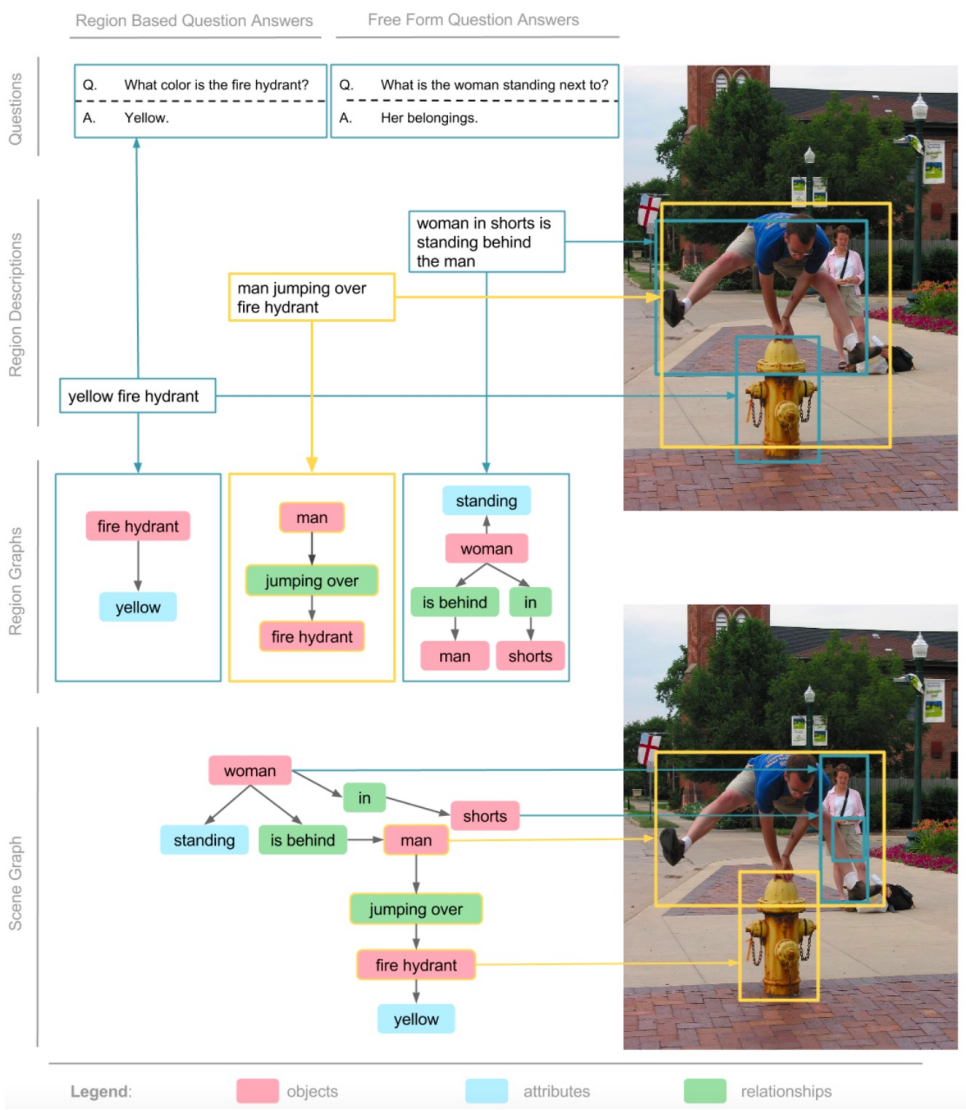
Keio University

Visual Genome [Krishna+ (Stanford Univ.), 17]

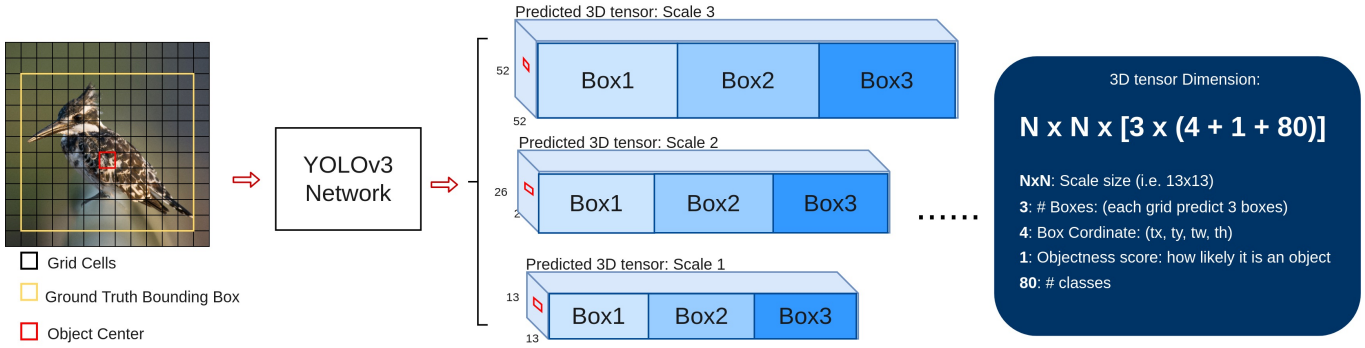
- 多くの高品質なラベルが付与された画像データセット
  - 10万8000枚以上
- 各画像に対し、平均50のregion description (1~16単語)
- 平均35のobjectで構成され、bboxで囲まれている
- 各objectはWordNetのsynset IDに正規化
  - man.n.03、person.n.01 など
    - personがmanの上位概念
- 平均26のattribute
  - objectには、0個以上のattribute
    - 色（例：yellow）、状態（例：standing） など
- 平均21のrelationship
  - jumping_over(man, fire hydrant)と表現
- 各Regionの有向グラフ表現 (Region Graph)
  - ノードは object、attribute、relationship
- 全体を表す1つのScene Graph
  - Region Graphを組み合わせたもの
- 2種類のQAペア
  - 画像全体に基づく自由形式のQAと、画像のある領域に基づいたQA



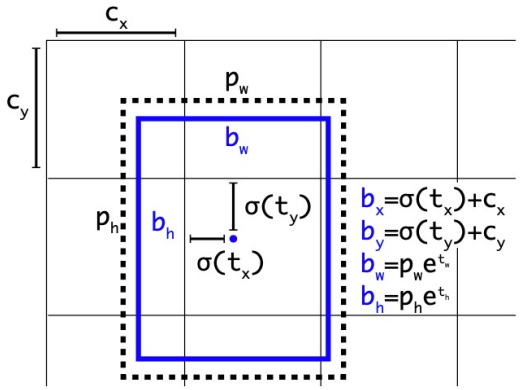
Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017).
Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123(1), 32-73.

YOLOv3 [Redmon+ (Univ. of Washington), 18]

- Darknet-53
  - ResNet-152と同程度の性能で2倍高速
- K-means法で学習用のbboxを作成
- 画像ごとに3つの3D tensorを予測
- ネットワークがt_x, t_y, t_w, t_hを予測
  - c_x, c_yはオフセット、
    p_w, p_hは作成したbboxの幅と高さ
    - 学習はground truthとの二乗誤差
  - bboxの選択にはロジスティック分類を使用
    - 各bboxのobjectness scoreを予測
- クラスはマルチラベル分類
  - ロジスティック分類を使用
    - 学習にはbinary cross-entropy loss
- v2は19層モデルに対しv3は53層



$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$



損失関数

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} \left[(t_x - \hat{t}_x)^2 + (t_y - \hat{t}_y)^2 + (t_w - \hat{t}_w)^2 + (t_h - \hat{t}_h)^2 \right]$$
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} \left[-log(\sigma(t_o)) + \sum_{k=1}^{C} BCE(\hat{y}_k, \sigma(s_k)) \right]$$
$$+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{noobj} \left[-log(1 - \sigma(t_o)) \right]$$

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Linear Transformers [Katharopoulos+ (Idiap Research Institute), ICML21]

- Attentionの計算に内積でなくカーネル関数を使用し、$O(n)$を達成
  - 長い配列の自己回帰予測では最大4000倍の速度
- i行目のAttentionV'_iを考える $A_l(x) = V' = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V.$ $V'_i = \frac{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j)}$ $\mathrm{sim}(q,k) = \exp\left(\frac{q^T k}{\sqrt{D}}\right)$
- $sim(q,k) = \phi(q)^T \phi(k)$ のようにq, kを分解

$$V'_i = \frac{\phi(Q_i)^T \boxed{\sum_{j=1}^{N} \phi(K_j) V_j^T}}{\phi(Q_i)^T \boxed{\sum_{j=1}^{N} \phi(K_j)}}$$ iに関与しないため一定

$$\sum_{j=1}^{n} \phi(k_j) \cdot v_j = A, \quad \sum_{j=1}^{n} \phi(k_j) = B \quad O(n) \qquad Attention(Q,K,V)_i = \frac{\phi(q_i)^T A}{\phi(q_i)^T B} \; O(1)$$

  - Softmaxは近似が難しいので $\phi(x) = \mathrm{elu}(x) + 1$ を使用
- 自己回帰モデルの学習
  - iがjから影響を受けるのは$j \leqq i$の場合のみ（後ろから影響を受けない）

$$V'_i = \frac{\sum_{j=1}^{i} \mathrm{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{i} \mathrm{sim}(Q_i, K_j)} \qquad V'_i = \frac{\phi(Q_i)^T S_i}{\phi(Q_i)^T Z_i} \qquad S_i = \sum_{j=1}^{i} \phi(K_j) V_j^T \quad Z_i = \sum_{j=1}^{i} \phi(K_j),$$

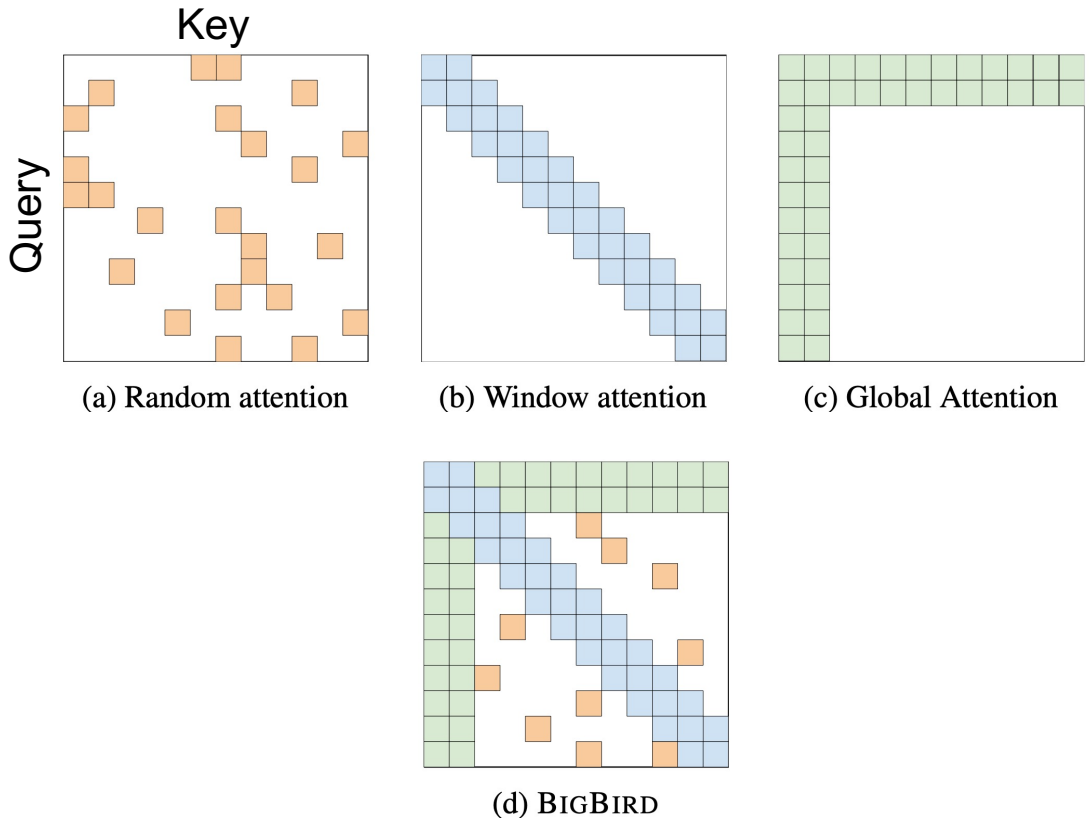
- S, Zを漸化式とみなすことでメモリ消費量を削減可能
  - f_lはFFN

$$s_0 = 0, \quad s_i = s_{i-1} + \phi(x_i W_K)(x_i W_V)^T, \qquad z_0 = 0, \quad z_i = z_{i-1} + \phi(x_i W_K), \qquad y_i = f_l\left(\frac{\phi(x_i W_Q)^T s_i}{\phi(x_i W_Q)^T z_i} + x_i\right).$$

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020, November). Transformers are rnns: Fast autoregressive transformers with linear attention. In International Conference on Machine Learning (pp. 5156-5165). PMLR.

BigBird [Zaheer+ (Google Research), NeurIPS20]

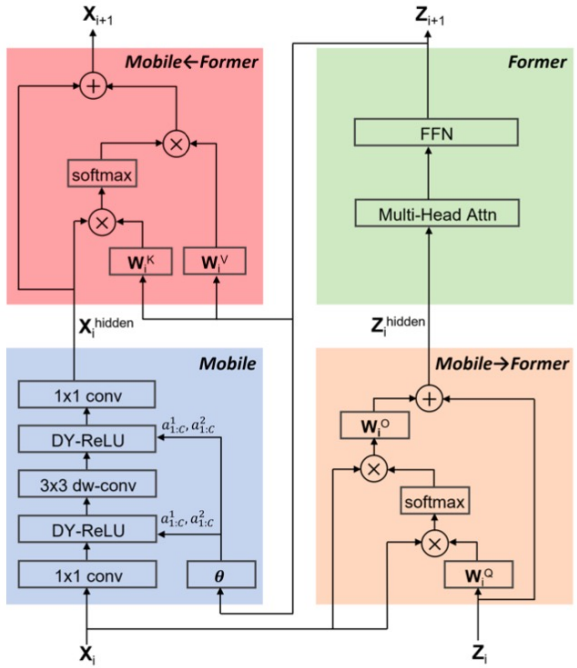
- Attentionにスパース性を導入し$O(n)$を達成
  - 長い文書を必要とするタスクについて高性能、文書要約タスクでSoTA
- QueryとKeyの乗算$O(n^2)$を近似
- Random Attention
  - 要素をランダムに選びattentionを計算
    - $r = 2$を選択、1行につき2要素をランダム
- Window Attention
  - 近傍のみattentionを計算
    - $w = 3$を選択、自分自身とその隣2つ
- Global Attention
  - 選択した要素と他の全要素のattentionを計算
    - $g = 2$を選択、行と列それぞれから2つ選択
- 上3つを足し合わせた箇所をAttention
  - 上は全て$O(n)$なので、Big Birdも$O(n)$

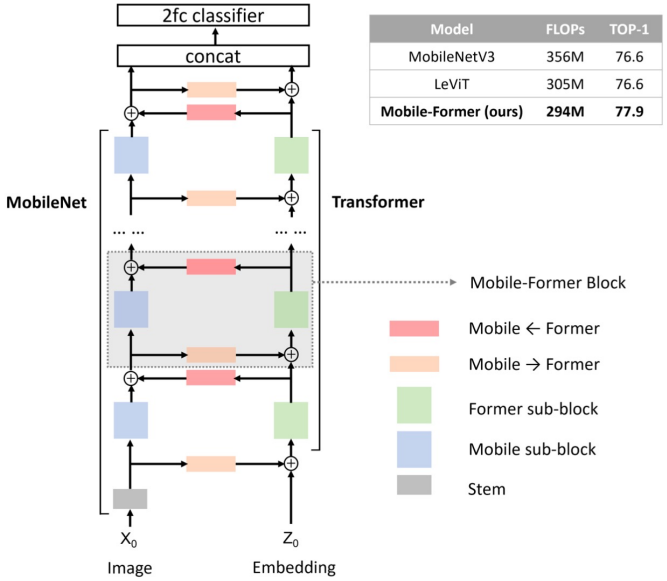


Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020, July). Big Bird: Transformers for Longer Sequences. In NeurIPS.

Mobile-Former [Chen+ (Microsoft), 21]

- ローカルな特徴量を扱うMobileNet+グローバルな特徴量を扱うTransformer
  - 計算コストを節約しつつ、精度を高めることが可能
- 2つの入力
  - 画像 $X_0 \in \mathbb{R}^{H \times W \times 3}$
  - 学習可能なトークン $Z_0 \in \mathbb{R}^{M \times d}$
    - ランダムに初期化　M:トークン数、M≦6
- Mobile → Former
  - $z^{out} = z + \left[Attention(z_h W_h^Q, x_h, x_h)\right]_{h=1:H} W^O$
- Former → Mobile
  - $x^{out} = x + \left[Attention(x_h, z_h W_h^K, z_h W_h^V)\right]_{h=1:H}$
    - H headのmulti-head attenのためx, zをsplit
      $x = [x_h]$ and $z = [z_h]$　$(1 \le h \le H)$
- 最終的なx, zをconcatし、線形層
  - ImageNet等で学習



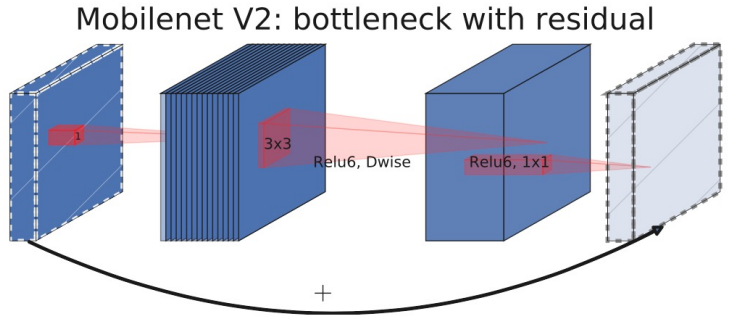


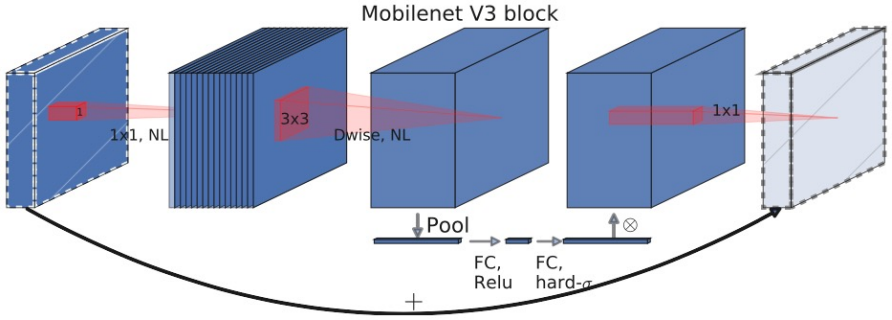
Model	FLOPs	TOP-1
MobileNetV3	356M	76.6
LeViT	305M	76.6
Mobile-Former (ours)	**294M**	**77.9**

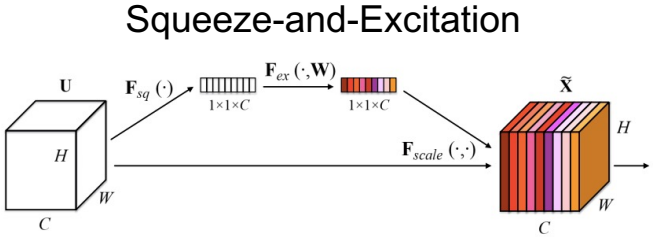
Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., & Liu, Z. (2021). Mobile-Former: Bridging MobileNet and Transformer. arXiv preprint arXiv:2108.05895.

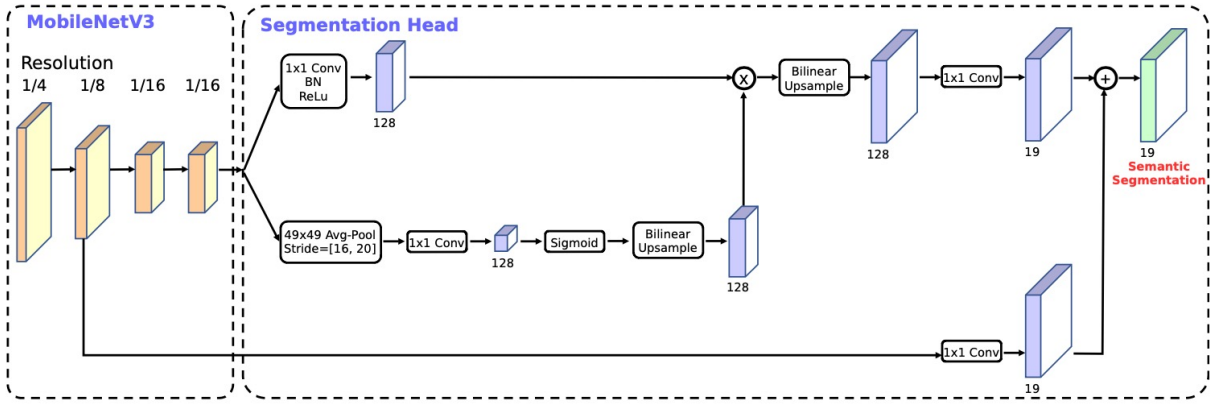
MobileNetV3 [Howard+ (Google AI), ICCV19]

- 携帯電話のCPUに合わせて調整された軽量のCNNモデル
  - object detectionやsemantic segmentationに適用
  - MobileNetV3-Large
    - V2からlatencyを20%削減し、ImageNet分類精度を3.2%向上
  - MobileNetV3-Small
    - V2と同じlatencyでImageNet分類精度を6.6%の向上
- Inverted Residual Block (V2)
  - 3x3 Depthwise Conv.を
    1x1(Pointwise) Conv.で挟む
- V3ではSqueeze-and-Excitationを導入
  - GAP→全結合層 したものを掛け合わせる
- Lite R-ASPP
  - 軽量なSegmentation Head
    - R-ASPPよりわずかに高速で、性能も向上







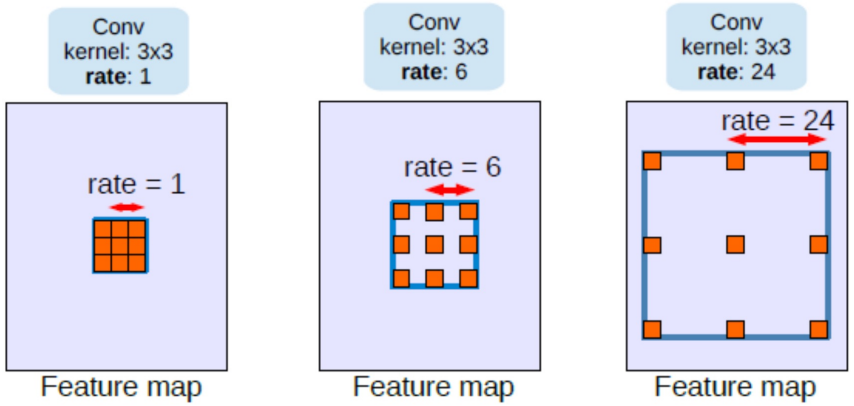


Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3.
In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314-1324).

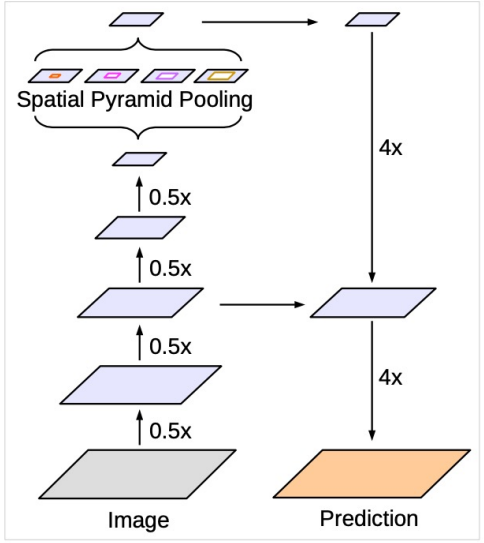
DeepLab v3+ [Chen+ (Google Inc.), ECCV18]

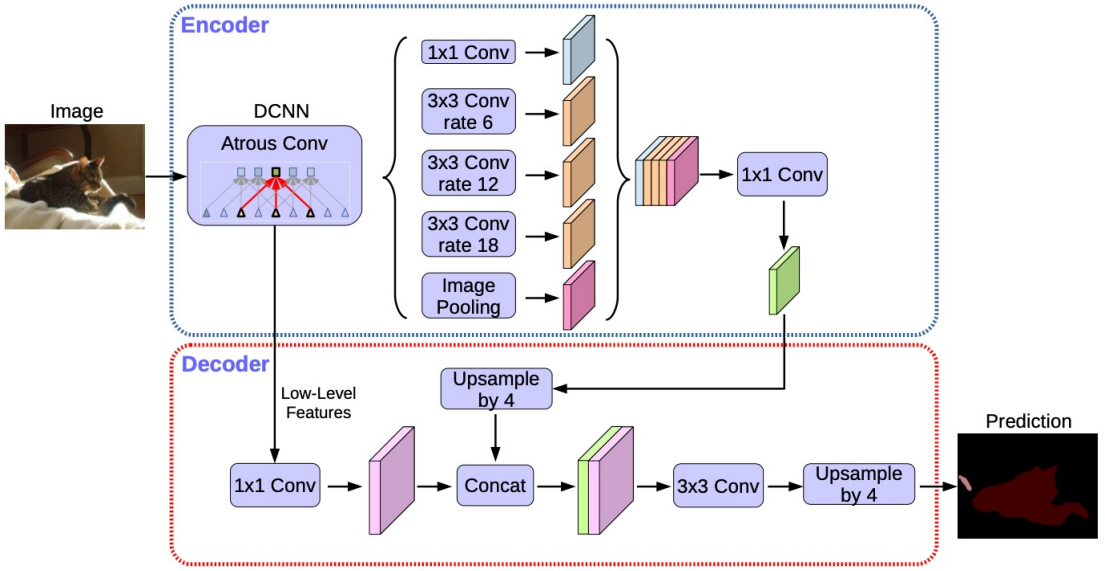
- Semantic SegmentationのためのEncoder-Decoderモジュール
- Atrous Convolution
  - マルチスケールの情報を捉える畳み込み

$$y[i] = \sum_{k} x[i + r \cdot k] w[k]$$



- Atrous Spatial Pyramid Pooling (ASPP)
  - 異なるスケールのAtrous Convolutionを特徴マップに適用
- Decoder
  - Low-level featuresを活用
  - bilinearly upsamplingを16倍から4倍に改善
    - 詳細なmapを作成





Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).