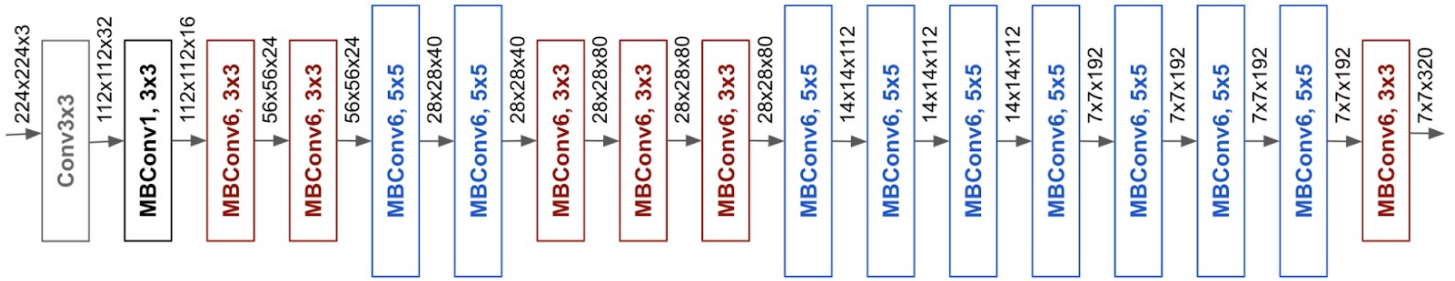
EfficientNet [Le+ (Google Research), ICML19]

- 深さ、幅、解像度のすべての次元を一様にスケーリング
  - 既存の最高のGPipeの精度を上回り、パラメータは8.4倍少なく、推論速度は6.1倍
- compound scaling method
  - ϕ で広さ、深さ、解像度を右式で決定
  - α = 1.2, β =1.1, γ = 1.15

$$\text{depth: } d = \alpha^\phi$$
$$\text{width: } w = \beta^\phi$$
$$\text{resolution: } r = \gamma^\phi$$
$$\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$
$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

- BaselineはMnasNetを使用
  - MBConv：Moble Inverted BottleneckにSEモジュールを追加



Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.

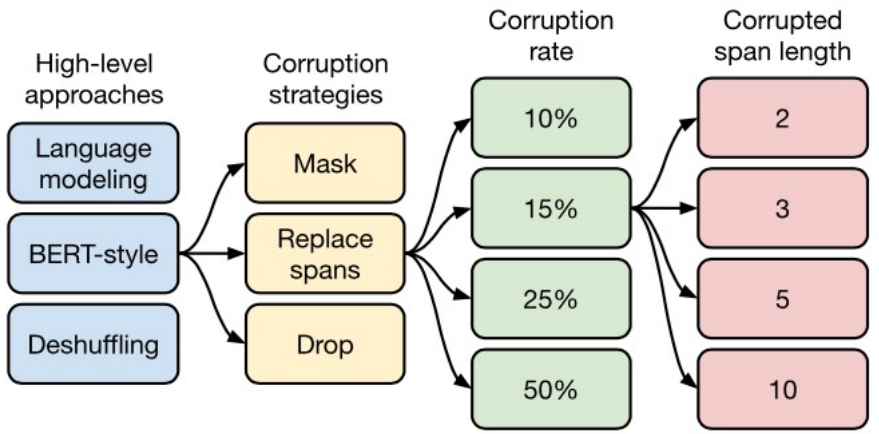
GPT-3 [Brown+ (OpenAI), NeurIPS20]

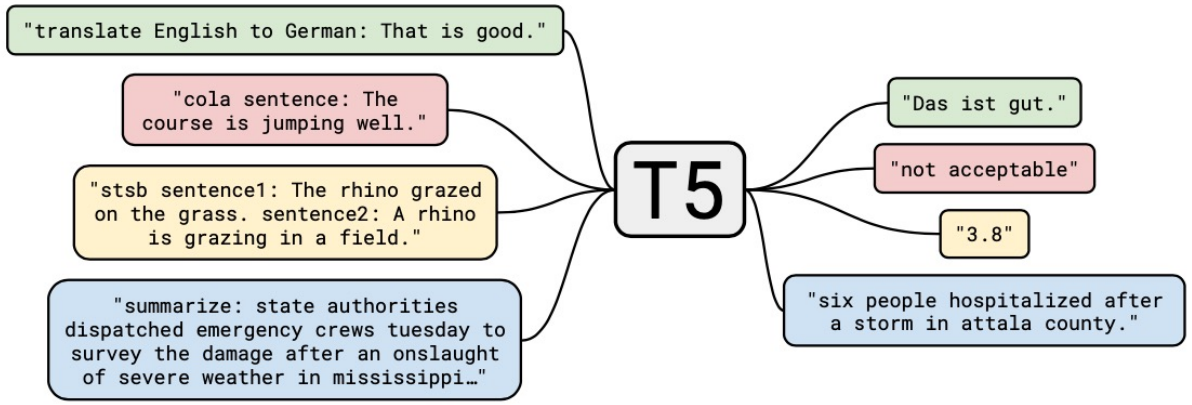
- Fine-tuningなしで様々なタスクを解くモデル
  - タスクに応じたデータセットが不要
  - 偏ったデータでfine-tuningすることによる汎化性能の低下を防ぐ
- 約45TBの大規模なテキストデータを約1750億個のパラメータを使用して学習
  - GPT-2と比較して使用データは1100倍以上、パラメータ数は117倍以上
- 各タスクにおいて、Fine-tuningしたモデルと同等の性能を達成

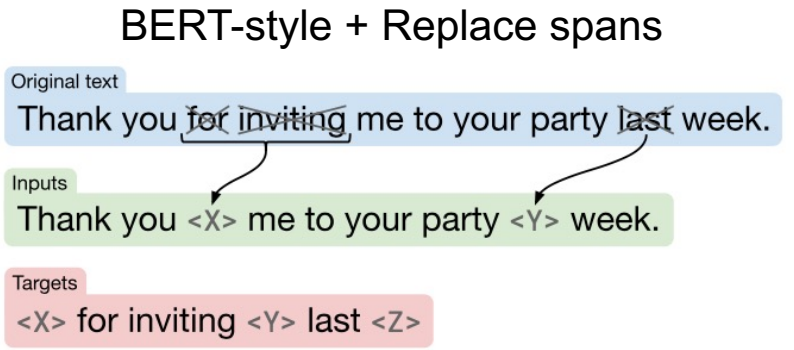
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

T5 [Raffel+ (Google Research), JMLR20]

- すべてのNLPタスクを “text-to-text” の形式として扱う
  - テキストを入力し、生成したテキストを出力
    - 入力テキストの接頭辞はハイパーパラメータ
  - 要約、質疑応答、テキスト分類など多くのタスクでSOTA
- Dataset：Colossal Clean Crawled Corpus(C4)
  - Common Crawl をクリーンアップ
  - そのまま使用する場合より精度が向上
- 事前学習について性能比較(右)
  - Corrupted Sapn lengthは大差なく、Baselineをそのまま使用



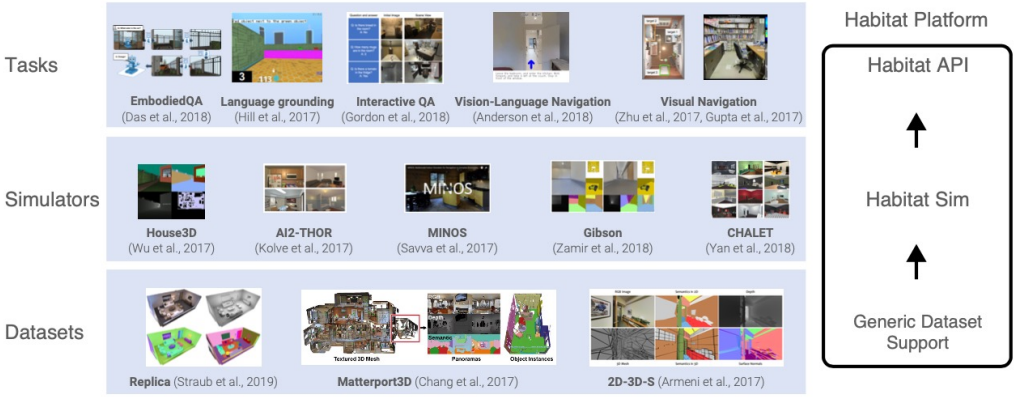




Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Habitat [Savva+ (Facebook AI Research), ICCV19]

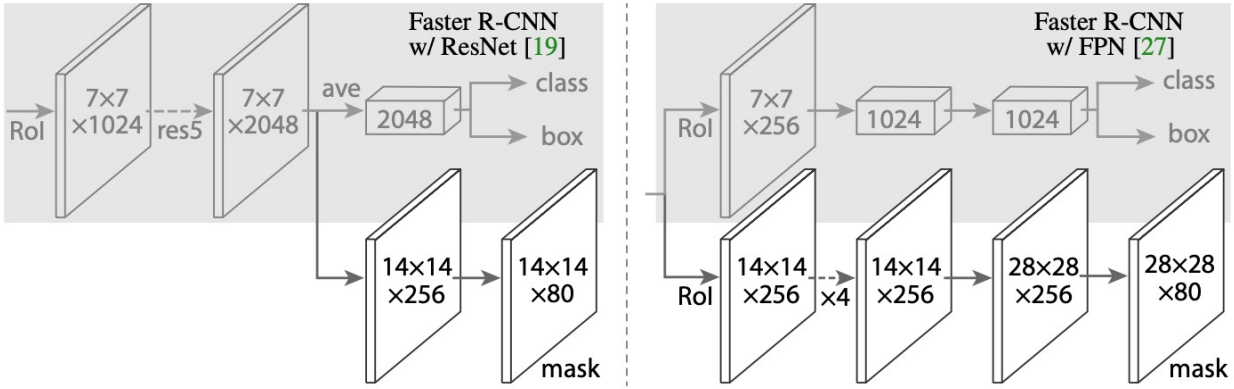
- Embodied AIのためのプラットフォーム
  - エージェントを高効率でフォトリアリスティックな3Dシミュレーションで育成
- Habitat-Sim
  - 柔軟で高性能な3次元シミュレータ
  - エージェント、センサー、汎用的な3Dデータセットを扱う
    - シングルスレッドで数千fps、シングルGPUではマルチプロセスで10,000fps以上を達成
- Habitat-API
  - 体現型AIアルゴリズムをエンドツーエンドで開発するためのモジュール式高レベルライブラリ
    - タスク（ナビゲーション、質問応答など）の定義、エージェントの設定とトレーニング等を行う



Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... & Batra, D. (2019).
Habitat: A platform for embodied ai research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9339-9347).

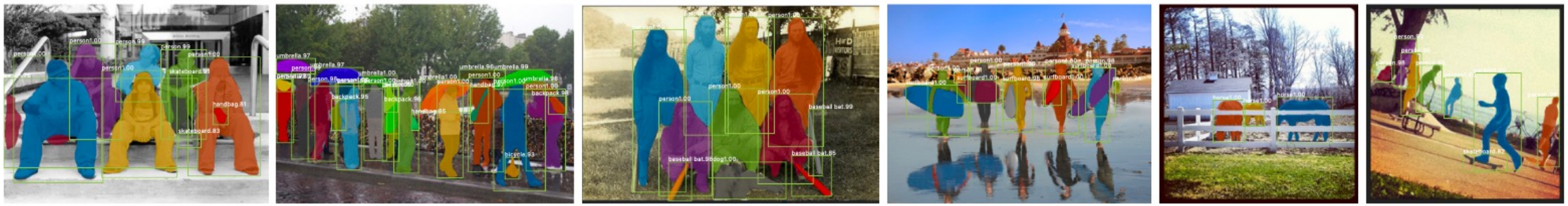
Mask R-CNN [He+ (Facebook AI Research), ICCV17]

- Object Detectionに加え、Instance Segmentatinに対応したモデル
  - Instance Segmentation タスクにおいてSoTAとなる37.1mAP
- Faster R-CNNにbranchを追加
  - それぞれのRoI(Region of Interest)に対してSegmentation Maskを予測



$$L = L_{cls} + L_{box} + L_{mask}$$

Binary Cross Entropy



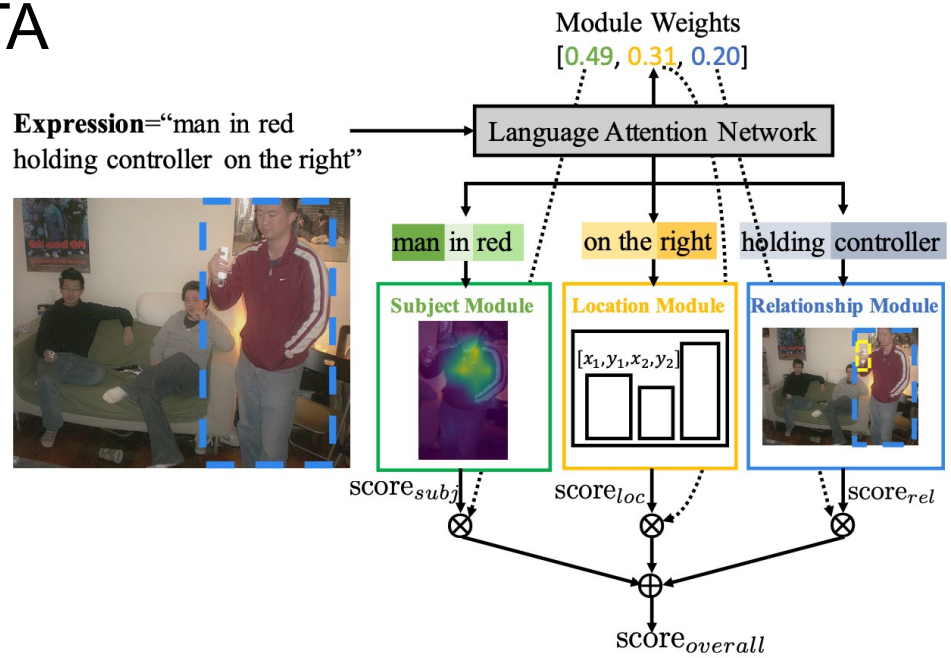
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

MAttNet [Yu+ (University of North Carolina at Chapel Hill), CVPR18]

- Referring expression comprehension においてSoTA
- 文章表現を3つに分解
  - subject appearance：全体像
  - Location：位置を示す
  - Relationship：物体同士の関係性
- FC層で3つのモジュールの重みに変換

$$e_t = \text{embedding}(u_t)$$
$$\vec{h}_t = \vec{\text{LSTM}}(e_t, \vec{h}_{t-1})$$
$$\overleftarrow{h}_t = \overleftarrow{\text{LSTM}}(e_t, \overleftarrow{h}_{t+1})$$
$$h_t = [\vec{h}_t, \overleftarrow{h}_t].$$

$$[w_{subj}, w_{loc}, w_{rel}] = \text{softmax}(W_m^T [h_0, h_T] + b_m)$$

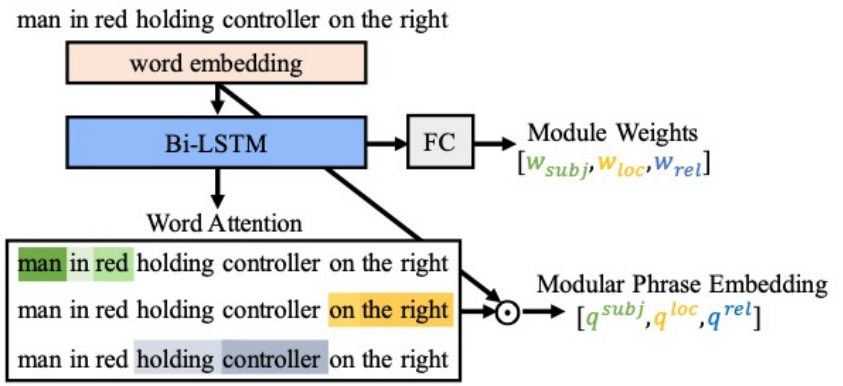


- Loss関数
  - o_iは候補オブジェクト、r_iはexpression

$$S(o_i|r) = w_{subj} S(o_i|q^{subj}) + w_{loc} S(o_i|q^{loc}) + w_{rel} S(o_i|q^{rel})$$

$$L_{rank} = \sum_i [\lambda_1 \max(0, \Delta + S(o_i|r_j) - S(o_i|r_i)) + \lambda_2 \max(0, \Delta + S(o_k|r_i) - S(o_i|r_i))]$$

$$L = L_{subj}^{attr} + L_{rank}$$



Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., & Berg, T. L. (2018). Mattnet: Modular attention network for referring expression comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1307-1315).

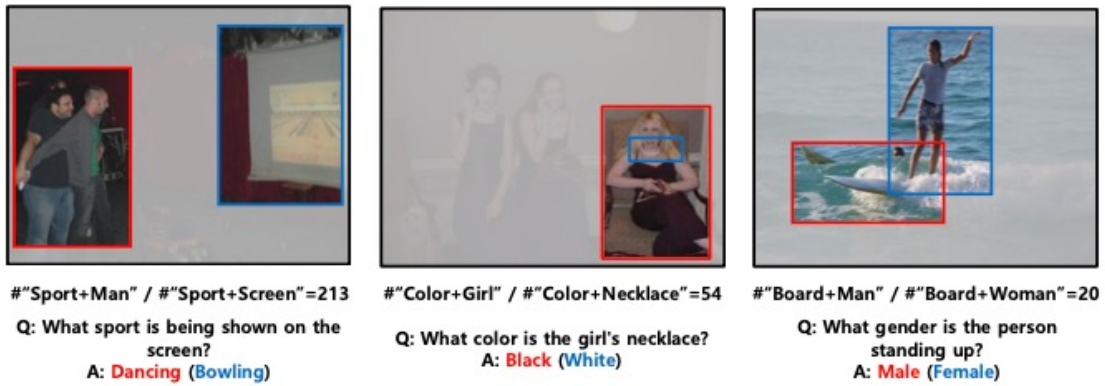
Causal Attention [Yang+ (Nanyang Technological University), CVPR21]

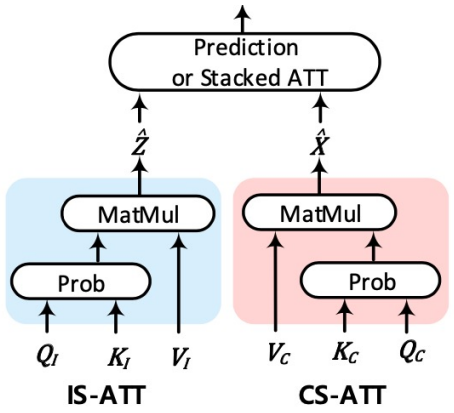
- 事前学習によるAttentionの悪化（右図）
  - トレーニングセットのバイアスによる
- 因果関係を考慮したCausal Attention(CATT)
  - 他のサンプルを用いてAttentionを算出
  - LXMERTに組み込むことで
    サイズの大きいUNITERと同等の精度
- In-Sample attention (IS-ATT)
  - K_I、V_Iは現在の入力サンプル

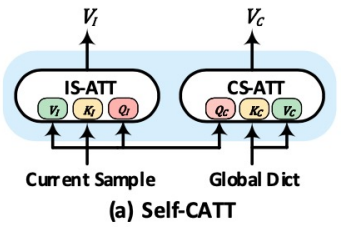
$$\textbf{Input:} \quad \boldsymbol{Q}_I, \boldsymbol{K}_I, \boldsymbol{V}_I,$$
$$\textbf{Prob:} \quad \boldsymbol{A}_I = \text{Softmax}(\boldsymbol{Q}_I{}^T \boldsymbol{K}_I)$$
$$\textbf{Ouput:} \quad \hat{\boldsymbol{Z}} = \boldsymbol{V}_I \boldsymbol{A}_I$$

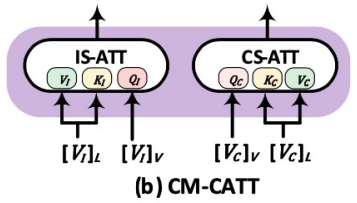
- Cross-Sample attention (CS-ATT)
  - K_C 、V_Cはトレーニングセットの他のサンプル

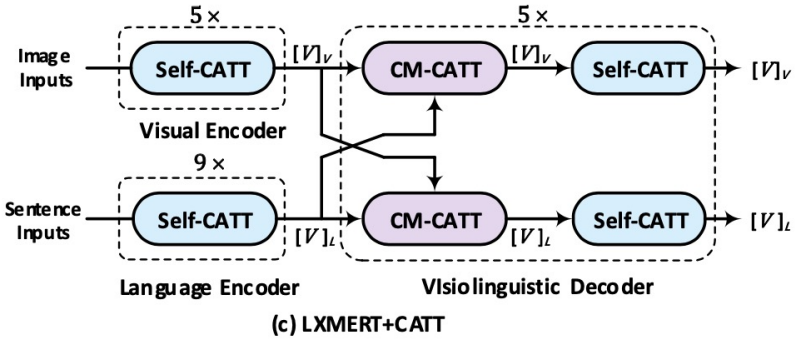
$$\textbf{Input:} \quad \boldsymbol{Q}_C, \boldsymbol{K}_C, \boldsymbol{V}_C,$$
$$\textbf{Prob:} \quad \boldsymbol{A}_C = \text{Softmax}(\boldsymbol{Q}_C{}^T \boldsymbol{K}_C)$$
$$\textbf{Ouput:} \quad \hat{\boldsymbol{X}} = \boldsymbol{V}_C \boldsymbol{A}_C$$











Yang, X., Zhang, H., Qi, G., & Cai, J. (2021). Causal attention for vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9847-9857).

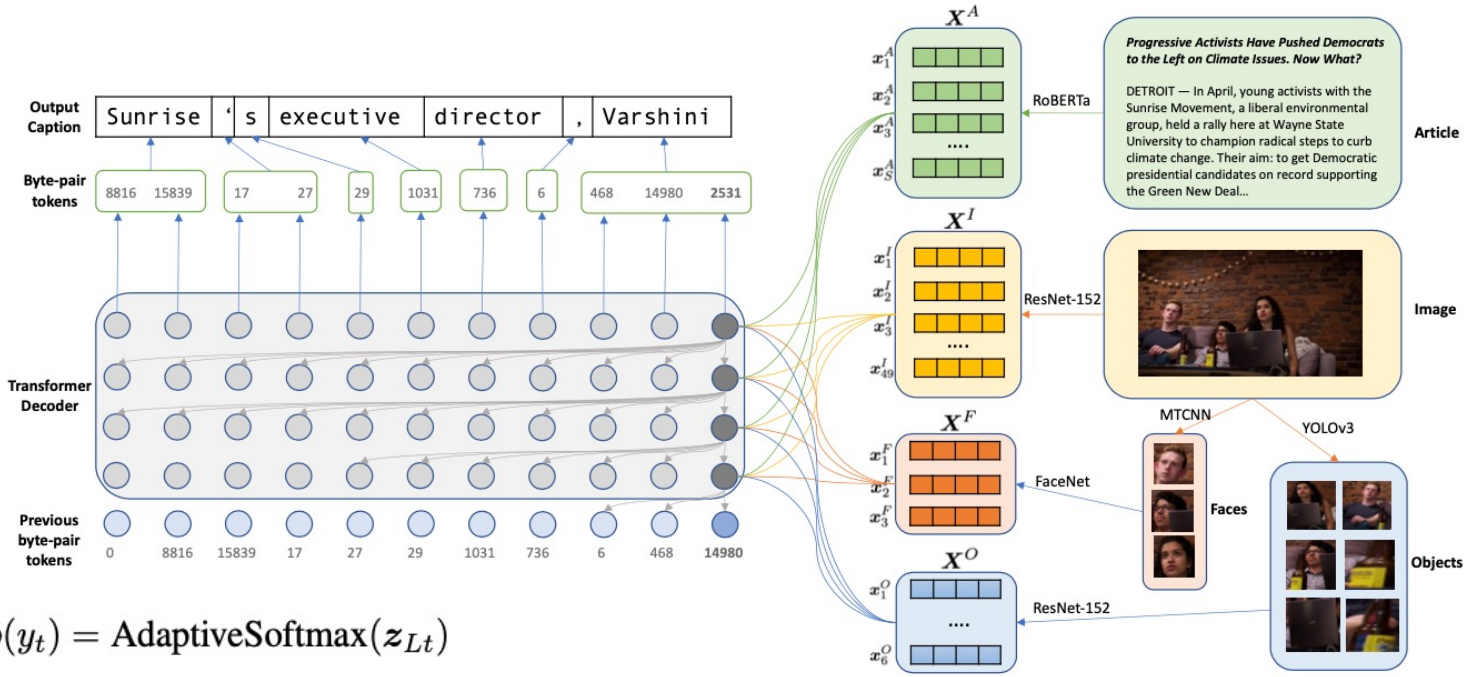
Transform and Tell [Tran+ (Australian National University), CVPR20]

- ニュース記事と埋め込まれた画像からキャプションを生成
  - BLEU-4：0.89→6.05、CIDEr：13.1→53.8
- Image Encoder：ResNet-152のプーリング層の前の出力
- Face Encoder：MTCNNでbboxを検出しFaceNetに通す、MFは顔の数
- Object Encoder：YOLOv3でbboxを検出しResNet-152に通す、MOはObject数
- Article Encoder：RoBERTaを使用、すべての層の加重和
- Decoder：3つの入力からキャプションを生成
  - 1：前ステップで生成したトークンのembedding
    $z_{0t} \in \mathbb{R}^{D^E}$ D^Eはhidden size
  - 2：前に生成された全トークンのembedding
    $Z_{0<t} = \{z_{00}, z_{01}, ..., z_{0t-1}\}$
  - 3：EncoderからのX^I、X^A、X^F、X^O
  - 入力はL個のtransformer blockに送られる

$$z_{1t} = \text{Block}_1(z_{0t}|Z_{0<t}, X^I, X^A, X^F, X^O)$$
$$z_{2t} = \text{Block}_2(z_{1t}|Z_{1<t}, X^I, X^A, X^F, X^O)$$
$$\dots$$
$$z_{Lt} = \text{Block}_L(z_{L-1t}|Z_{L-1<t}, X^I, X^A, X^F, X^O)$$

  - Z_{Lt}からt番目トークンの生成確率p(y_t)を推定：$p(y_t) = \text{AdaptiveSoftmax}(z_{Lt})$

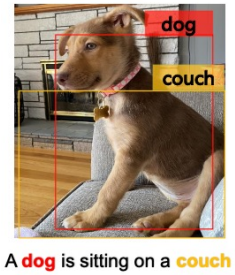
$$X^I = \{x_i^I \in \mathbb{R}^{D^I}\}_{i=1}^{M^I}, \text{where } D^I = 2048 \text{ and } M^I = 49$$
$$X^F = \{x_i^F \in \mathbb{R}^{D^F}\}_{i=1}^{M^F}, \text{where } D^F = 512$$
$$X^O = \{x_i^O \in \mathbb{R}^{D^O}\}_{i=1}^{M^O}, \text{where } D^O = 2048$$
$$X^A = \{x_i^A \in \mathbb{R}^{D^T}\}_{i=1}^{M^T}, \text{where } D^T = 1024$$

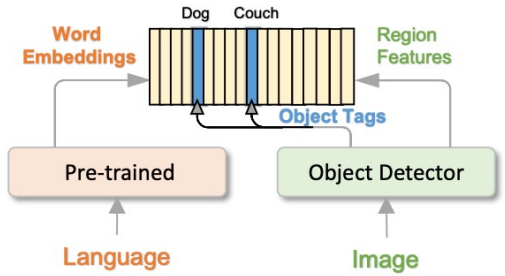


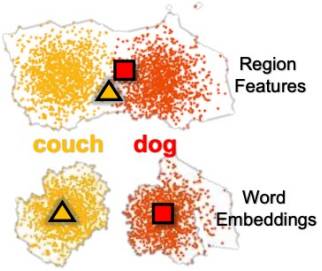
Tran, A., Mathews, A., & Xie, L. (2020). Transform and tell: Entity-aware news image captioning.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13035-13045).

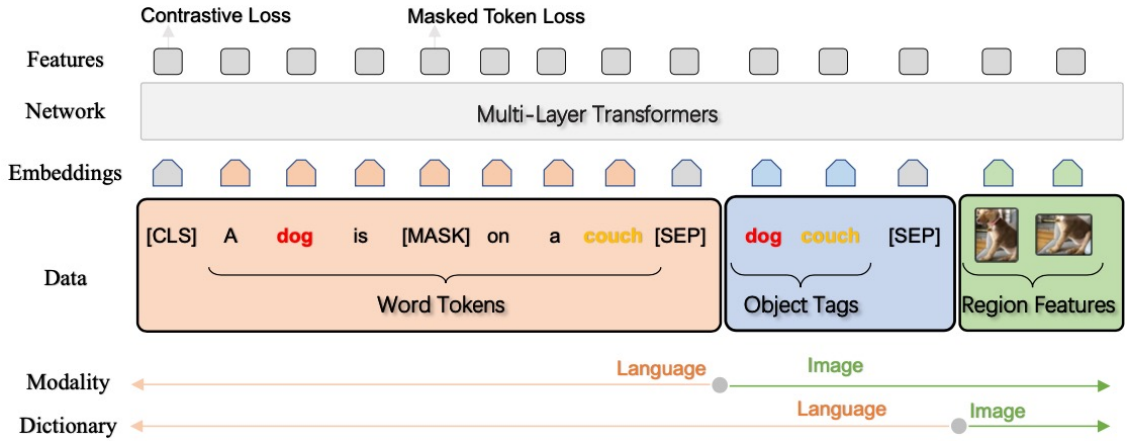
Oscar [Li+ (Microsoft Corporation), ECCV20]

- 画像-テキスト表現を学習するVLP手法
  - 複数のV+Lベンチマークで SoTA
- アンカーポイントとしてオブジェクトタグを使用
- Word Tokens (w)
  - テキストの単語埋め込みベクトル
- Object Tag (q)
  - 画像から検出されたオブジェクトタグの単語埋め込みベクトル
- Region Features (v)
  - 画像の領域ベクトルの集合
- 事前学習タスク
  - Masked Token Loss $\mathcal{L}_{\mathrm{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i}, \boldsymbol{v})$
    - w, qをマスクし、復元するよう学習
  - Contrastive Loss $\mathcal{L}_{\mathrm{C}} = -\mathbb{E}_{(\boldsymbol{h}',\boldsymbol{w})\sim\mathcal{D}} \log p(y|f(\boldsymbol{h}', \boldsymbol{w}))$
    - qを50%の確率で異なるタグに置き換え、適切なタグかどうか予測









Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020, August). Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision (pp. 121-137). Springer, Cham.