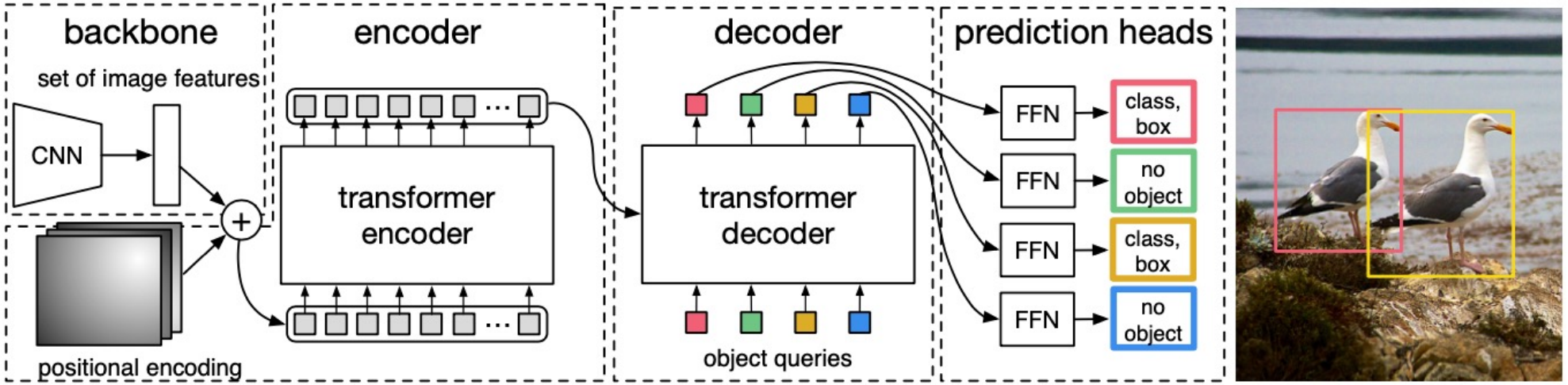
DETR [Carion+ (Facebook AI), ECCV20]

- 一度の処理で画像内のすべての物体を検出しクラスを予測
- 画像内の物体の関係性を利用
- Transformerを用いることで並列にデコード可能
  - 既存研究は殆どが自己回帰モデル



- 小さい物体の検出や学習時間が課題

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European Conference on Computer Vision (pp. 213-229). Springer, Cham.

Keio University

RoBERTa [Liu+ (Facebook AI), ICLR20]

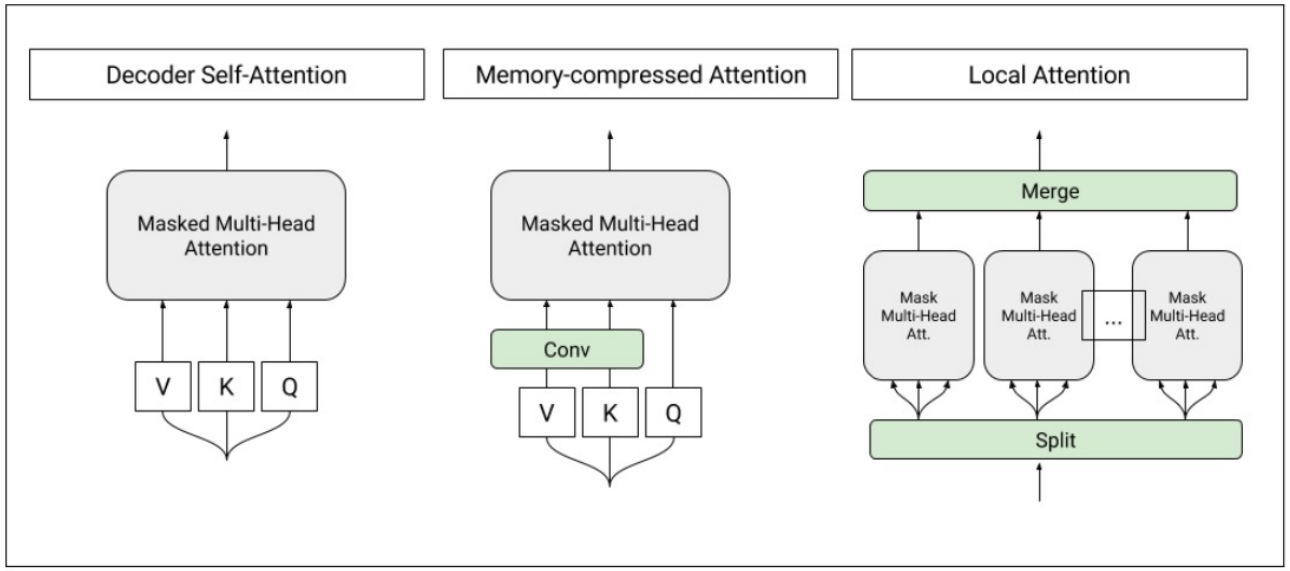
- BERTの事前学習を改良し、性能を大幅に向上
- 大きなバッチで学習
  - 256(BERT)→2k
- 多くのデータで、長い時間(10万～50万ステップ)で学習
  - Dataset：16GB(BERT)→160GB
- 学習の度にマスクの箇所を変更
  - BERTでは最初に一度決定したマスクを使用
- より長い文章を入力として学習
  - BERTでは短いセグメントを多く使用
- Next Sentence Prediction(NSP)は効果がない
  - BERTで用いられていた事前学習タスク
  - 除くことで精度向上

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Memory Compressed Transformer [J. Liu+ (Google Brain), ICLR18]

- 文章の要約タスクに関して長シーケンスをTransformerで扱う
- Memory-compressed Attention
  - Valye、Keyに対して畳み込みを行い数を減らす
- Local Attention
  - シーケンスを同じような長さのブロックに分割し、各ブロックで独立してattentionを行う

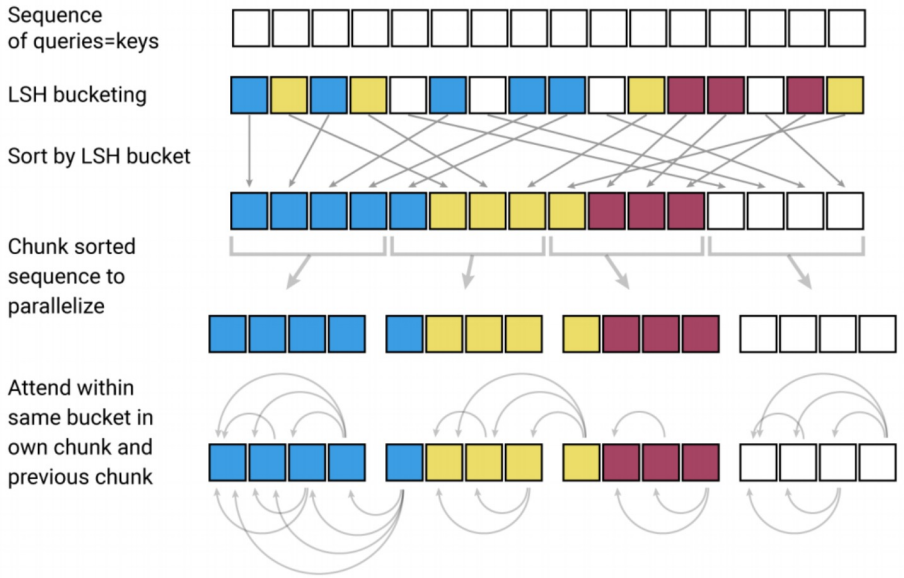


Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198.

Reformer [Nikita Kitaev+ (U.C. Berkeley & Google Research), 20]

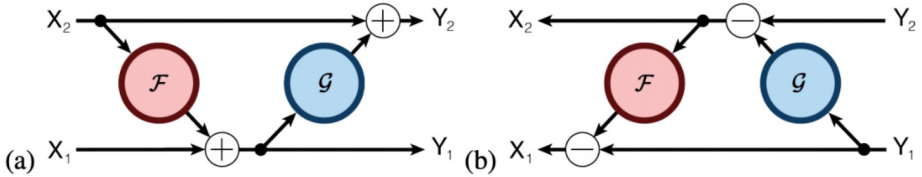
- 長いシーケンスLで高速に動作し、メモリ効率が高いTransformer
  - locality-sensitive hashing (LSH) によりO(L^2) からO(L logL)に削減

- LSH Attention
  - 元々の位置より手前側、同一のバケット内、1つ前までのチャンク、という条件を満たす場合のみAttention



- Reversible layers
  - ResNetに類似した構造
  - 出力のactivationから入力のactivationを簡単な計算で復元
    - レイヤー数Nに比例して増える activationを保存しておくメモリを削減

$$y_1 = x_1 + \mathcal{F}(x_2)$$
$$y_2 = x_2 + \mathcal{G}(y_1)$$

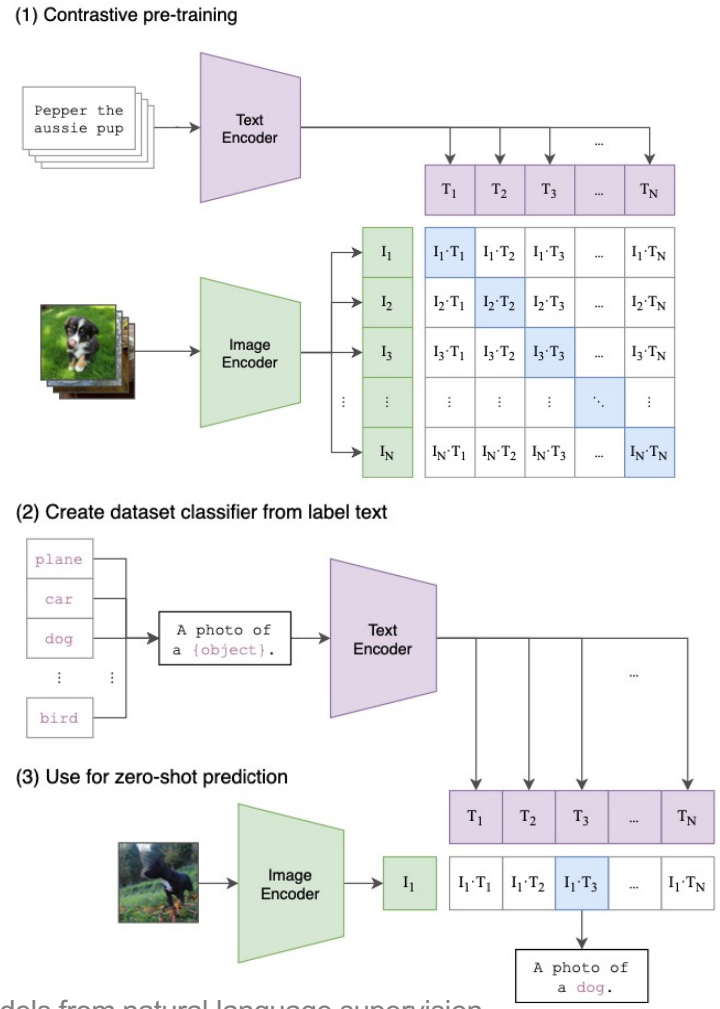


$$x_2 = y_2 - \mathcal{G}(y_1)$$
$$x_1 = y_1 - \mathcal{F}(x_2)$$

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451.

CLIP [Radford+ (OpenAI), 21]

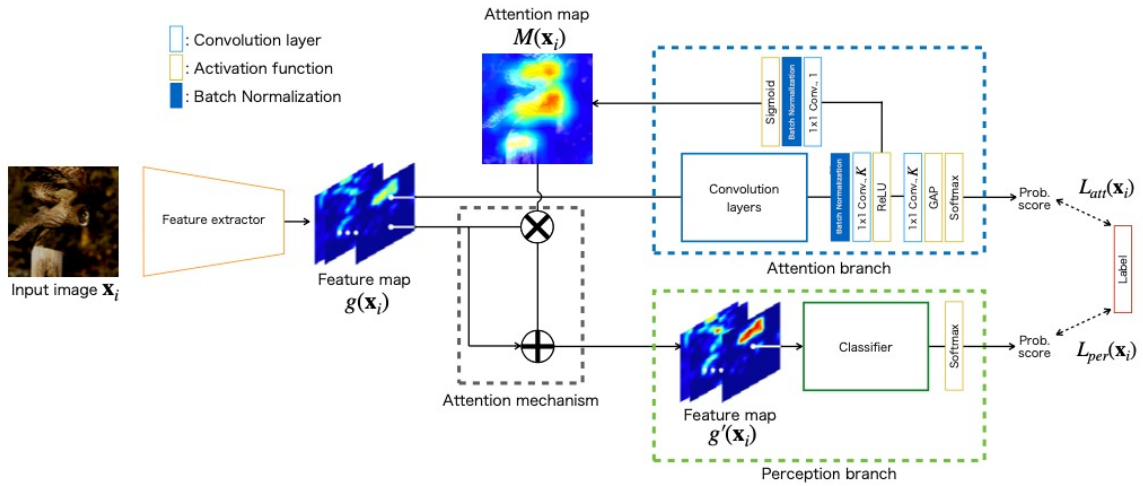
- インターネット上の画像とテキストのペアで事前学習することでラベル付けのコストを削減
  - 4億組の学習データ
- Contrastive Pre-Training
  - 正解ラベルが複数の候補のうちどれかを選択する事前学習
- Prompt Engineering
  - 推論時、正解ラベルをA photo of {label}とすることで精度を向上
    - 事前学習時のラベルと近づく
- Zero-shotで高精度の分類が可能
  - 教師あり学習を行ったResNet-50と互角の性能
  - 衛星画像や腫瘍検出等の複雑なタスクでは精度が下がる



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Attention Branch Network [Fukui+ (Chubu University), CVPR21]

- Attentionの可視化と精度向上を同時に実現
- CNNをFeature extractorとPerception branchに分割
- Feature extractorからAttention branchへ分岐
- Attention mapとFeature extractorの出力を合わせてPercepton branchに入力
  - 両方のブランチでの損失を使用して学習可能



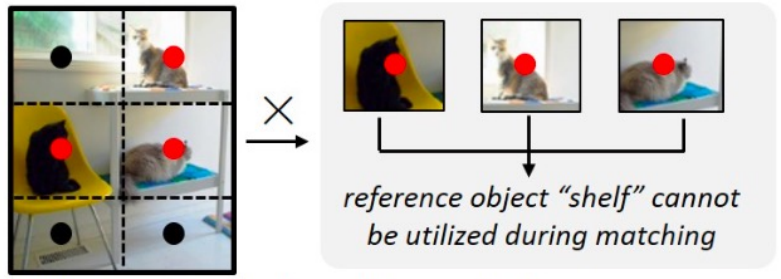
- 今後は、学習過程にラベルを含まない強化学習にもABNを適用する予定

Fukui, H., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10705-10714).

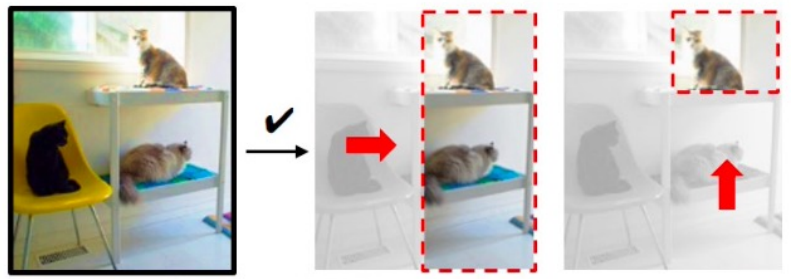
Iterative Shrinking with RL [Sun+ (University of Liverpool), CVPR21]

- 強化学習ベースの物体検出
  - Proposal-freeのreferring expression grounding task
- 予測のboxを縮小する操作を強化学習のactionとして定義
  - エージェントが縮小方向を決定
- 報酬
  - $$r_i = \begin{cases} 0 & IoU < 0.3 \ or \ \Delta_{IoU} \leqslant 0 \\ 1 & 0.3 \leqslant IoU < 0.5 \ and \ 0 < \Delta_{IoU} \\ 10 & 0.5 \leqslant IoU \ and \ 0 < \Delta_{IoU} \end{cases}$$
- 画像内のすべてのオブジェクトを総合的に判断
  - 参照表現の理解に優位

「the cat above the shelf」



(a). feature-point level matching method

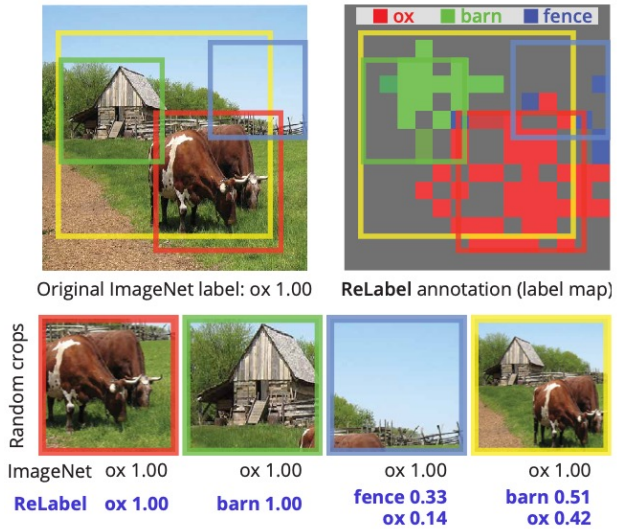


(b). proposed iterative shrinking method

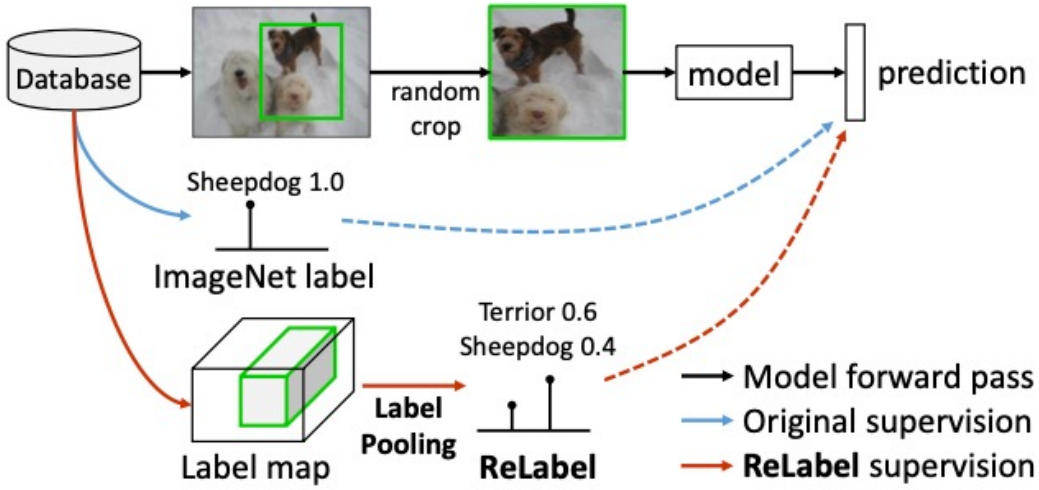
Sun, M., Xiao, J., & Lim, E. G. (2021). Iterative Shrinking for Referring Expression Grounding Using Deep Reinforcement Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14060-14069).

ReLabel [Yun+ (NAVER AI Lab), CVPR21]

- Image Net に多くのラベルノイズ
  - 多くのサンプルに複数のクラスが含まれている
  - Random crop時には正解が画像に含まれていないことも
- マルチラベルタスクとして画像ごとにアノテーション
- 膨大なデータセットで学習した識別器でlabel mapを作成



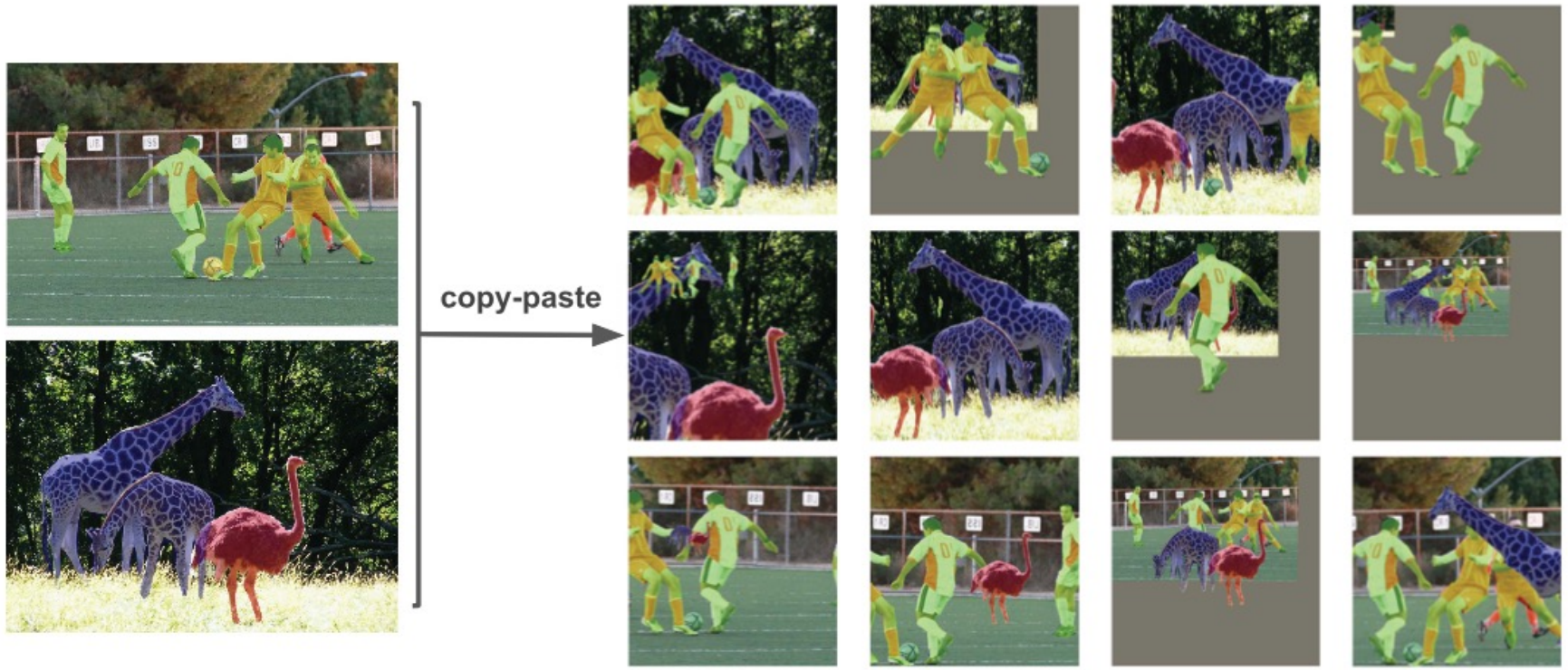
- LabelPooling
  - ランダムに切り抜かれた箇所に関して対応したラベルを学習に使用



Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., & Chun, S. (2021). Re-labeling imagenet: from single to multi-labels, from global to localized labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2340-2350).

Copy-Paste [Golnaz Ghiasi+ (Google Research), CVPR21]

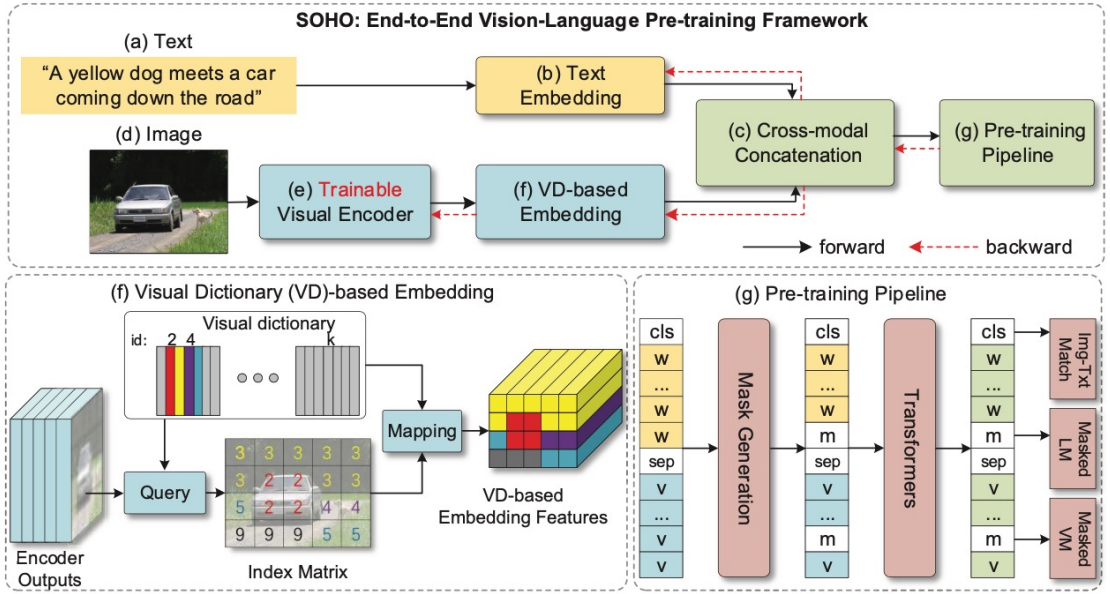
- 画像のスケーリング、コピー＆ペーストによる非常にシンプルなData Augmentation
- 片方の画像からランダムにオブジェクトのサブセットを選択し、もう片方の画像に貼り付け
- Instance SegmentationやObject Detectionで従来のデータ拡張より優れた結果



Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., ... & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2918-2928).

SOHO [Huang+ (University of Science and Technology Beijing), CVPR21]

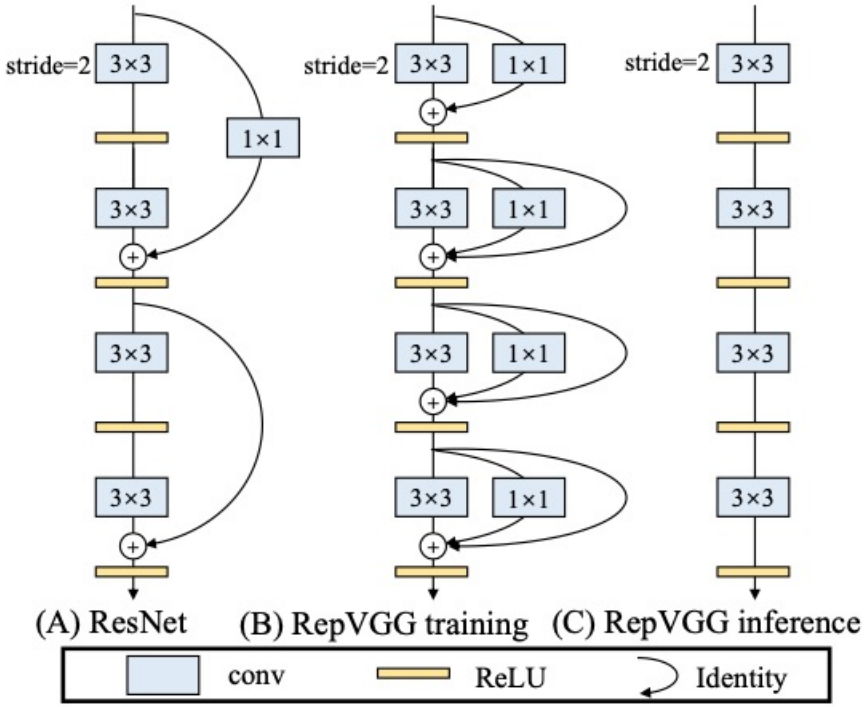
- Seeing Out of tHe bOx
- VL事前学習において自然言語のペアとして画像全体を入力
  - bounding boxを抽出しないため推論を10倍以上高速化
  - 検出した物体と周りの領域との関係を学習可能
- Visual Dictionary
  - 視覚特徴をトークン化するために、類似した視覚特徴量を同じ特徴量に集約

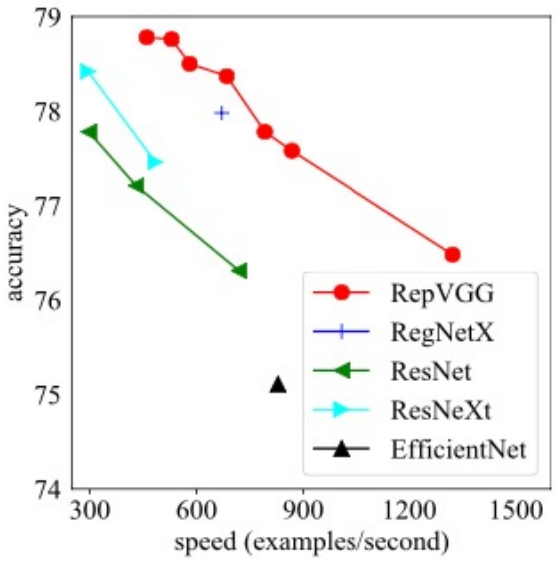


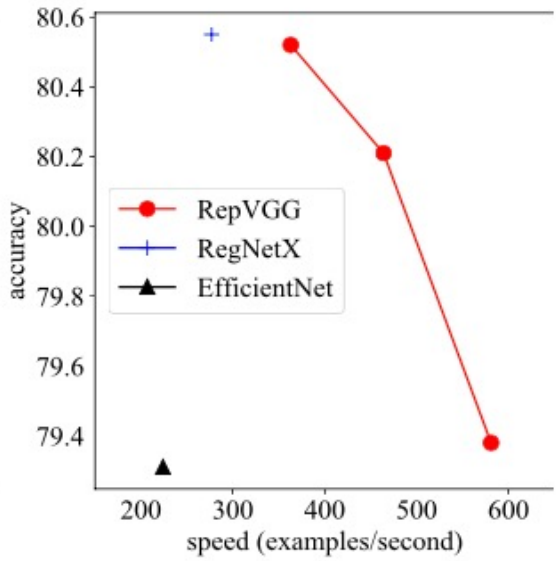
Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., & Fu, J. (2021). Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12976-12985).

RepVGG [Ding+ (BNRist), CVPR21]

- 学習時は3x3conv、1x1conv、skip-connectionの3つのブランチを使用
- 推論時は分岐を使用しないシンプルなネットワーク
- ResNet-50よりも83%、ResNet-101よりも101%高速に動作し、高精度
  - EfficientNetのような最先端のモデルと比較しても、精度と速度の良好なトレードオフを実現



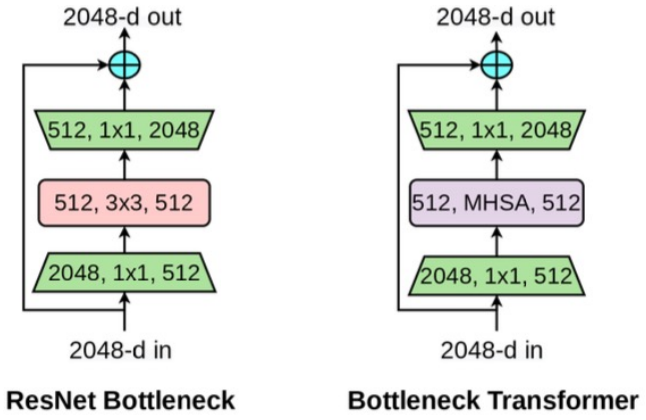




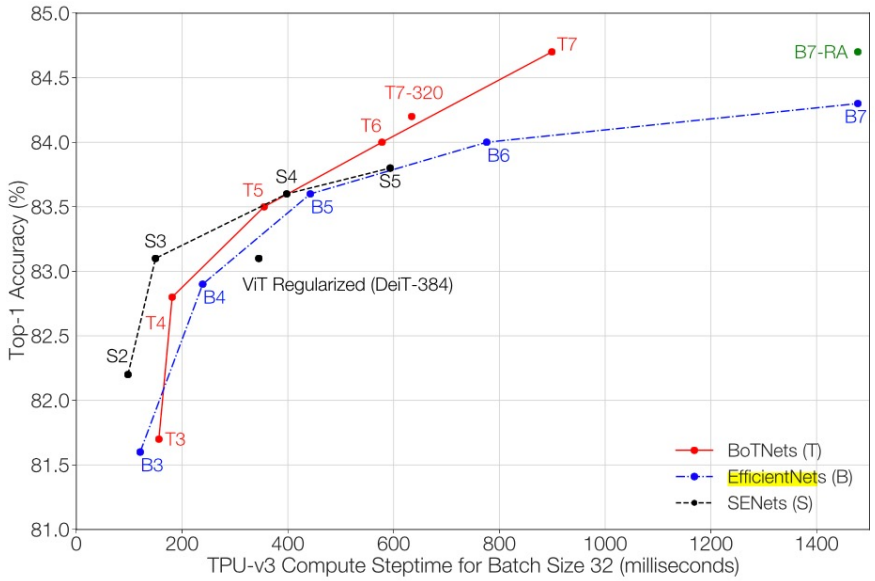
Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13733-13742).

BoTNet [Tsung-Yi Lin+ (Google Research), CVPR21]

- ResNetの最後の3つの3×3畳み込みをmulti-head self-attentionに置き換え
  - パラメータを削減し、レイテンシのオーバーヘッドを最小限に抑える
  - ハイパーパラメータ等の変更が不要
- Instance segmentation、Object detectionでベースラインを大幅に改善
- EfficientNetと同等の精度を達成し、また1.64倍高速に動作



stage	output	ResNet-50	**BoTNet-50**
c1	512 × 512	7×7, 64, stride 2	7×7, 64, stride 2
c2	256 × 256	3×3 max pool, stride 2	3×3 max pool, stride 2
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
c3	128 × 128	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
c4	64 × 64	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
c5	32 × 32	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ \text{MHSA}, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
# params.		**25.5**×10^6	**20.8**×10^6
M.Adds		**85.4**×10^9	**102.98**×10^9
TPU steptime		**786.5** ms	**1032.66** ms



Srinivas, A., Lin, T. Y., Parmar, N., Shlens, J., Abbeel, P., & Vaswani, A. (2021). Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16519-16529)