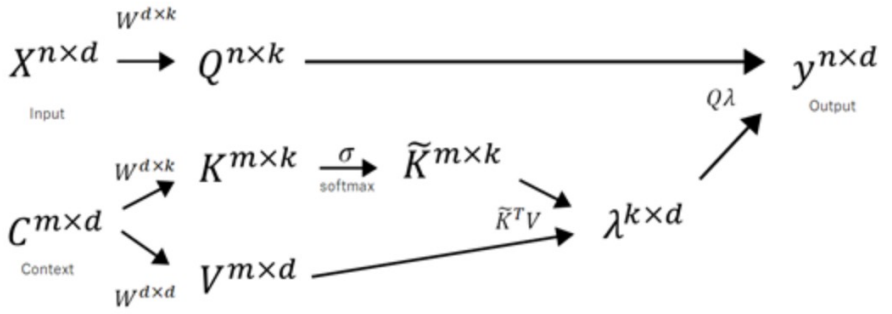
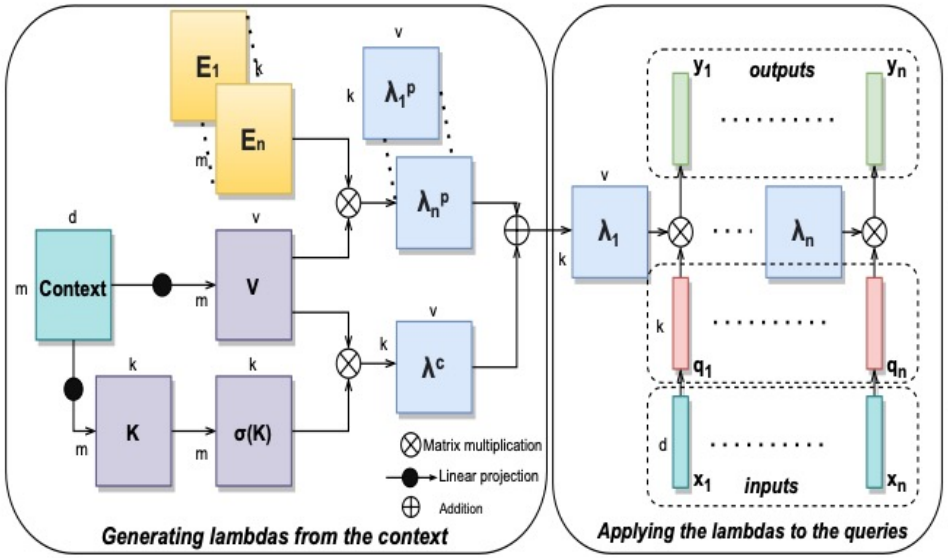
Lambda Networks [Bello (Google Research), ICLR2021]

- Lambda Layer
  - Attentionを用いずに、広範囲の関係性を考慮
  - Contextを線形関数Lambdaに変換し、対応するQueryに直接適用
    - KeyとValueの積を先に取る
    - Attentionは $(QK^T)V$



- Self-Attentionと比べて少ないメモリで高い精度

Layer	Space Complexity	Memory (GB)	Throughput	top-1
Global self-attention	$\Theta(blhn^2)$	120	OOM	OOM
Axial self-attention	$\Theta(blhn\sqrt{n})$	4.8	960 ex/s	77.5
Local self-attention (7x7)	$\Theta(blhnm)$	-	440 ex/s	77.4
Lambda layer	$\Theta(lkn^2)$	1.9	1160ex/s	**78.4**
Lambda layer ($\|k\|$=8)	$\Theta(lkn^2)$	0.95	**1640** ex/s	77.9
Lambda layer (shared embeddings)	$\Theta(kn^2)$	**0.63**	1210 ex/s	78.0
Lambda convolution (7x7)	$\Theta(lknm)$	-	1100 ex/s	78.1



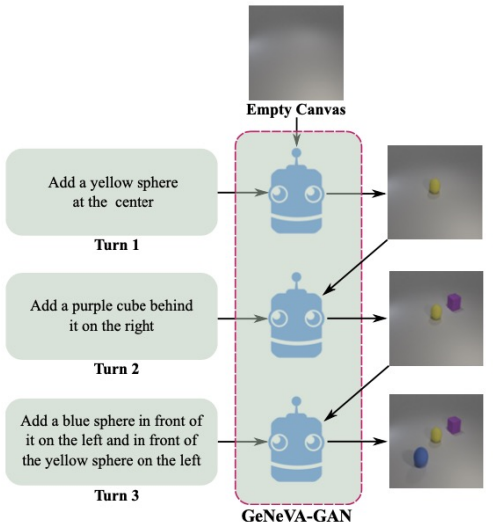
Name	Description
$\|k\|, \|v\|$	query, value depth
$\boldsymbol{X} \in \mathbb{R}^{\|n\|\times d}$	inputs
$\boldsymbol{C} \in \mathbb{R}^{\|m\|\times d}$	context
$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q \in \mathbb{R}^{\|n\|\times\|k\|}$	queries
$\boldsymbol{K} = \boldsymbol{C}\boldsymbol{W}_K \in \mathbb{R}^{\|m\|\times\|k\|}$	keys
$\boldsymbol{V} = \boldsymbol{C}\boldsymbol{W}_V \in \mathbb{R}^{\|m\|\times\|v\|}$	values
$\sigma(\boldsymbol{K}) = \text{softmax}(\boldsymbol{K}, \text{axis}{=}m)$	normalized keys
$\boldsymbol{E}_n \in \mathbb{R}^{\|m\|\times\|k\|}$	relative position embeddings
$\boldsymbol{\lambda}^c = \bar{\boldsymbol{K}}^T\boldsymbol{V} \in \mathbb{R}^{\|k\|\times\|v\|}$	*content* lambda
$\boldsymbol{\lambda}_n^p = \boldsymbol{E}_n^T\boldsymbol{V} \in \mathbb{R}^{\|k\|\times\|v\|}$	*position* lambdas
$\boldsymbol{\lambda}_n = \boldsymbol{\lambda}^c + \boldsymbol{\lambda}_n^p \in \mathbb{R}^{\|k\|\times\|v\|}$	lambdas

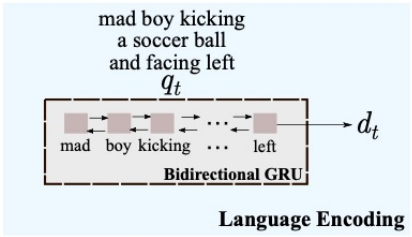
$$\boldsymbol{y}_n = \boldsymbol{\lambda}_n^T \boldsymbol{q}_n = (\boldsymbol{\lambda}^c + \boldsymbol{\lambda}_n^p)^T \boldsymbol{q}_n \in \mathbb{R}^{|v|}.$$

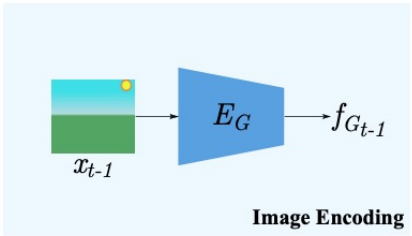
Bello, I. (2021). Lambdanetworks: Modeling long-range interactions without attention. *arXiv preprint arXiv:2102.08602*.

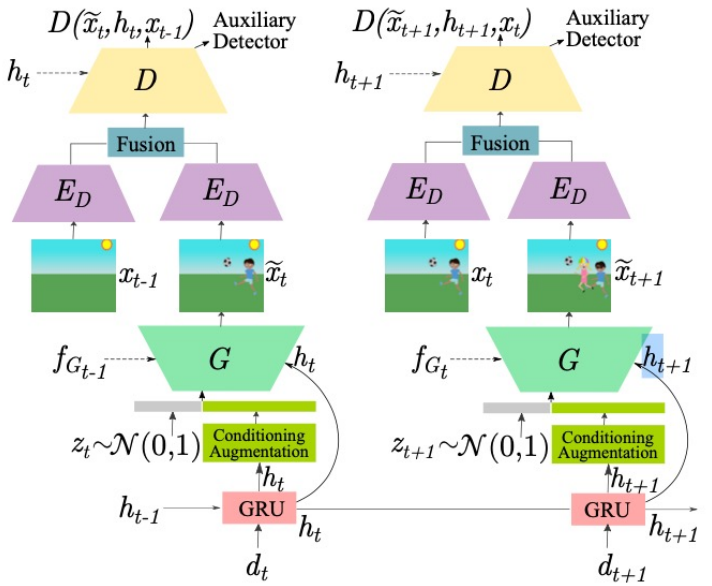
GeNeVa-GAN [El-Nouby+ (University of Guelph), ICCV19]

- 現在まで生成された出力と過去の全ての命令、両方を考慮する回帰GANモデル
- 対象：GeNeVa task
  - テキストを入力とし、全てのテキストを満たすような画像を出力
- Gated Recurrent Unit (GRU)：命令特徴量d_tと前の条件h_{t-1}から条件h_tを出力
- Generator (G)：条件h_tと時刻t-1の教師画像特徴量f_{Gt-1}から画像$\tilde{x}_t$を生成
- Image Encoder (E_D)：t-1の教師画像x_{t-1}と生成画像$\tilde{x}_t$をエンコードし、出力をFusion
- Discriminator (D)：実画像と生成画像を区別する他、4つのlossを算出
  - $$L_{D_{\text{real}}} = -\mathbb{E}_{(x_t,c_t)\sim p_{\text{data}(0:T)}}[\min(0,-1+D(x_t,c_t))]$$
    $$L_{D_{\text{wrong}}} = -\mathbb{E}_{(x_t,\hat{c}_t)\sim p_{\text{data}(0:T)}}[\min(0,-1-D(x_t,\hat{c}_t))]$$
    $$L_{D_{\text{fake}}} = -\mathbb{E}_{z_t\sim\mathcal{N},c_t\sim p_{\text{data}(0:T)}}[\min(0,-1-D(G(z_t,\tilde{c}_t),c_t))]$$
    $$\tilde{c}_t = \{h_t, f_{G_{t-1}}\} \text{ and } c_t = \{h_t, x_{t-1}\}.$$
    - $\hat{C}_t$は誤った命令によるもの
  - 全ての物体が存在するか検出
    $$L_{\text{aux}} = \sum_{i=0}^{N} -(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$









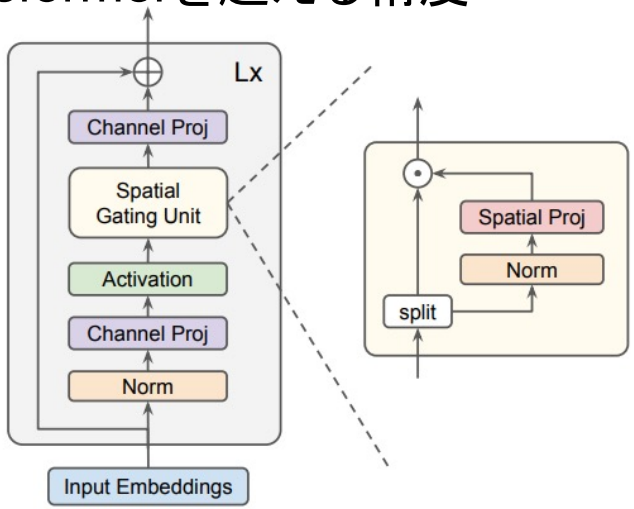
El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L. E., Kahou, S. E., ... & Taylor, G. W. (2019).
Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10304-10312).

gMLP [Liu+ (Google Research), 21]

- Attentionを使用しないMLPベースのモデルでTransformerと同等の性能
- Spatial Gating Unit
  - 入力をチャネル方向に2分割
  - 片方を線形投射し2つを乗算

$$s(Z) = Z_1 \odot f_{W,b}(Z_2) \quad f_{W,b}(Z) = WZ + b$$

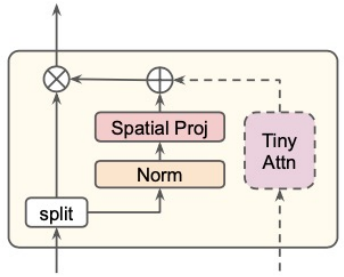
- aMLP
  - 複数文を扱うタスクではSelf-Attentionが重要
  - サイズ64、シングルヘッドの小さなAttention blockを追加
  - Transformerを超える精度



Pseudo-code for the gMLP block

```
def gmlp_block(x, d_model, d_ffn):
  shortcut = x
  x = norm(x, axis="channel")
  x = proj(x, d_ffn, axis="channel")
  x = gelu(x)
  x = spatial_gating_unit(x)
  x = proj(x, d_model, axis="channel")
  return x + shortcut

def spatial_gating_unit(x):
  u, v = split(x, axis="channel")
  v = norm(v, axis="channel")
  n = get_dim(v, axis="spatial")
  v = proj(v, n, axis="spatial", init_bias=1)
  return u * v
```



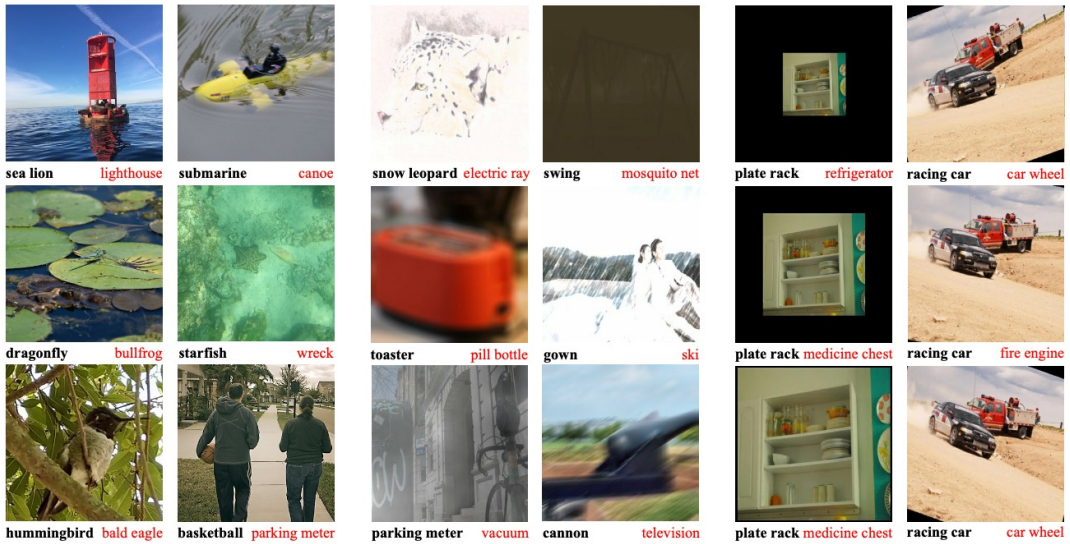
Pseudo-code for the tiny attention module

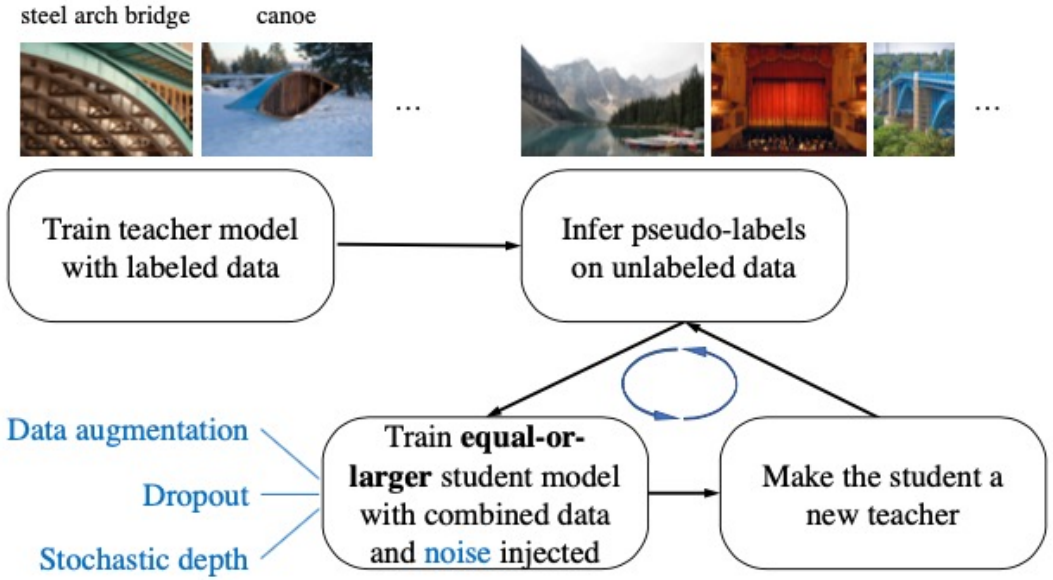
```
def tiny_attn(x, d_out, d_attn=64):
  qkv = proj(x, 3 * d_attn, axis="channel")
  q, k, v = split(qkv, 3, axis="channel")
  w = einsum("bnd,bmd->bnm", q, k)
  a = softmax(w * rsqrt(d_attn))
  x = einsum("bnm,bmd->bnd", a, v)
  return proj(x, d_out, axis="channel")
```

Liu, H., Dai, Z., So, D. R., & Le, Q. V. (2021). Pay Attention to MLPs. arXiv preprint arXiv:2105.08050.

Noisy Student [Xie+ (Google Research), CVPR20]

- Noisy Student Training
  - Self-trainingにおいて、Studentの学習時に大きくノイズをかける
    - ノイズ：RandAugment、Dropout、Stochastic Depth
  - 学習したStudentを新たにTeacherとし、より大きなStudentを学習することを繰り返す
- Noisy Studentを使用したEfficientNet
  - ImageNetのTop-1精度は88.4 %でSoTAを達成
  - 加えて高いロバスト性を示す



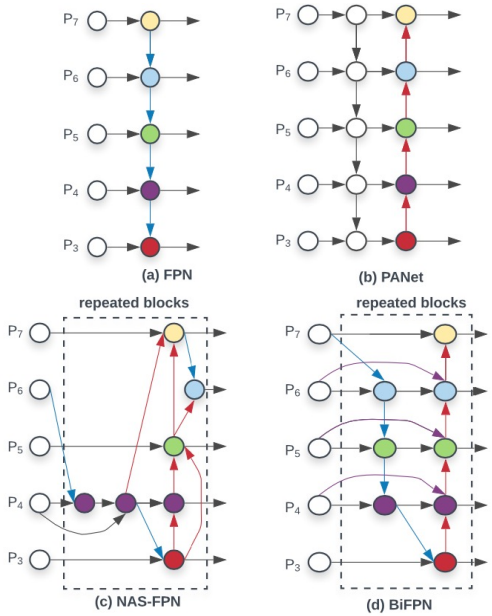


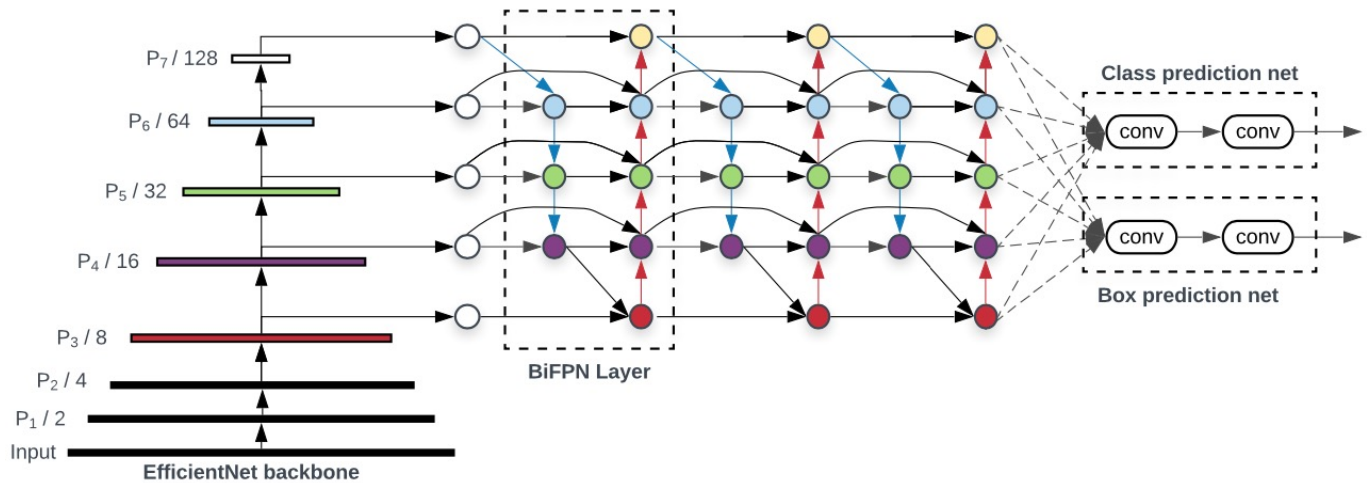
Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10687-10698).

EfficientDet [Xie+ (Google Research), CVPR20]

- 既存手法よりも少ないパラメータ数で同等以上の性能の物体検出モデル
  - EfficientDet-D7はCOCO Datasetで52.2 APを達成しSoTA
- BiFPN
  - 複数解像度の特徴量をトップダウンとボトムアップの両方で繰り返し流す
- Compaund Scaling
  - 一つのパラメータϕによってネットワーク容量の調整を代表

	Input size R_{input}	Backbone Network	BiFPN #channels W_{bifpn}	BiFPN #layers D_{bifpn}	Box/class #layers D_{class}
D0 ($\phi = 0$)	512	B0	64	3	3
D1 ($\phi = 1$)	640	B1	88	4	3
D2 ($\phi = 2$)	768	B2	112	5	3
D3 ($\phi = 3$)	896	B3	160	6	4
D4 ($\phi = 4$)	1024	B4	224	7	4
D5 ($\phi = 5$)	1280	B5	288	7	4
D6 ($\phi = 6$)	1280	B6	384	8	5
D6 ($\phi = 7$)	1536	B6	384	8	5



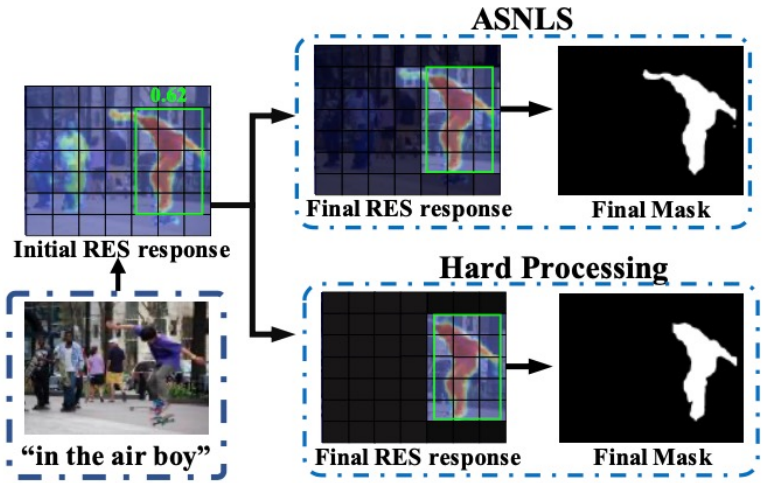


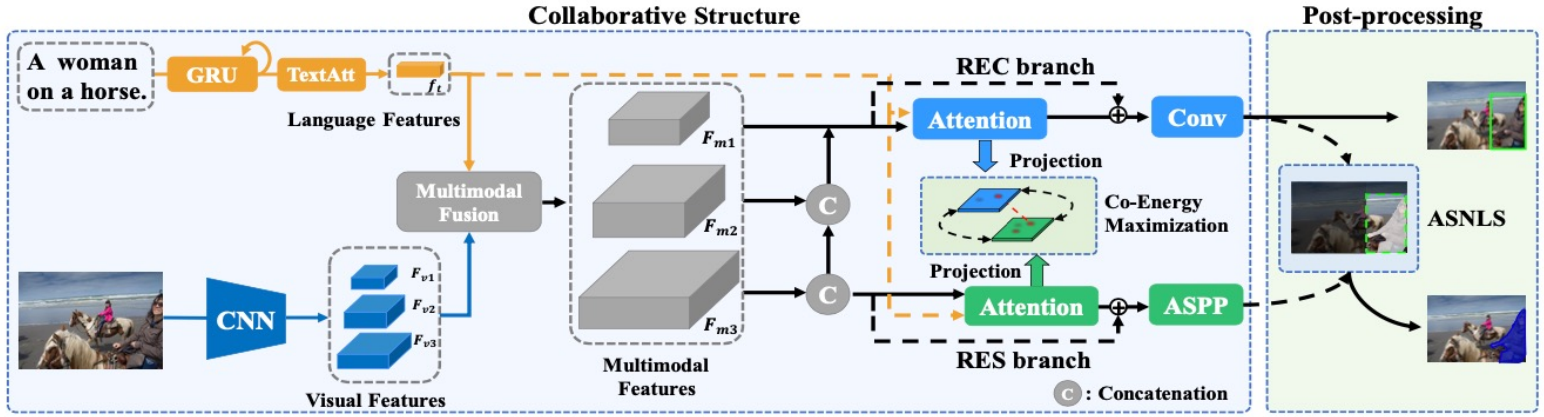
Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10781-10790)

Keio University

Multi-task Collaborative Network(MCN) [Luo+ (Xiamen University), CVPR20]

- Referring Expression Comprehension(REC) と Segmentation(RES) を同時に扱うマルチタスク学習
  - REC：7.13%、RES：11.50%の大幅な性能向上
    - RECは6倍の推論速度
- Consistency Energy Maximization(CEM)
  - RECとRESが同じ領域に注目するように最適化するロス
- Adaptive Soft Non-Lacated Supression(ASNLS)
  - RECで予測したbounding boxに基づいて、RESの領域を制御
  - bounding box外の領域もある程度考慮



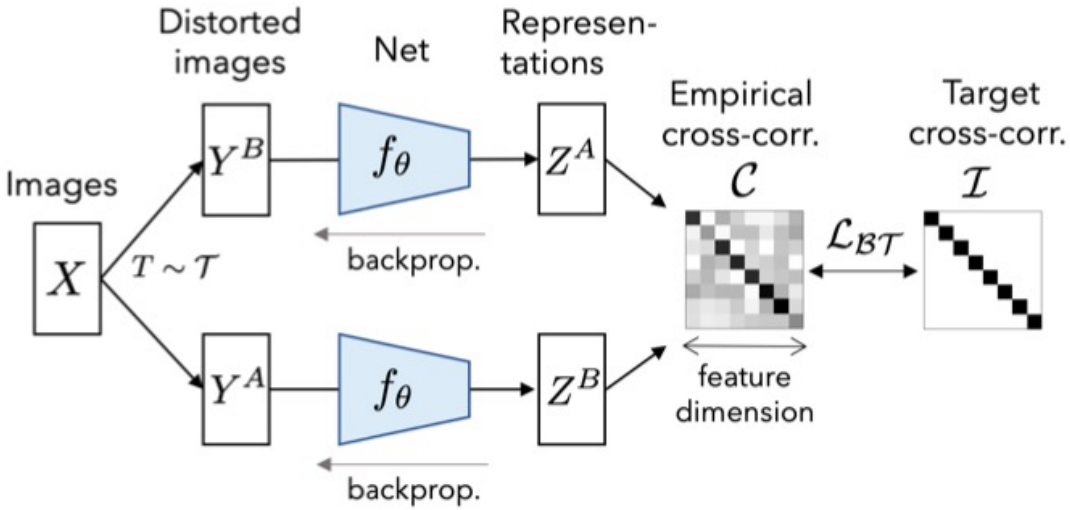


Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., & Ji, R. (2020). Multi-task collaborative network for joint referring expression comprehension and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10034-10043).

Barlow Twins [LeCun+ (Facebook AI Research), ICML21]

- 分類、物体検出のための新しい自己教師あり学習アルゴリズム
  - バッチサイズが小さい場合でも機能し、出力ベクトルを高次元化することが可能
- 入力に対し2パターン、ノイズを適用
  - クロップ、リサイズ + 水平方向の反転、グレー変換、ガウスぼかし、ソラリゼーション
- f_θ：ResNet-50 + projection (linear*3)
- 2つのベクトルから相互相関行列を算出
  - -1~1の正方行列

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$



- 単位行列に近くなるように学習
  - 入力の歪みに対して表現を不変 + 出力ユニット間の冗長性を減らす

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}}$$

1に近づける　　0に近づける

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230.