南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Query2Label: A Simple Transformer Way to Multi-Label Classification

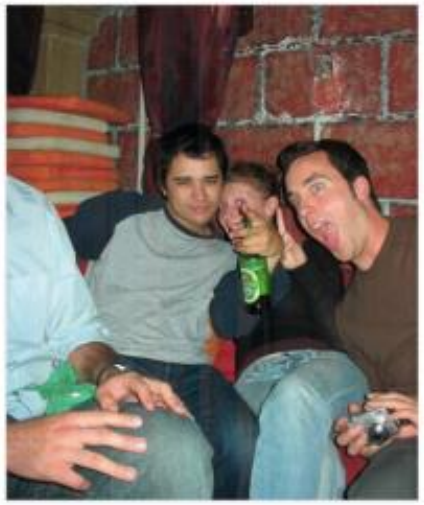
Shilong Liu[1,2], Lei Zhang[2], Xiao Yang[1], Hang Su[1], Jun Zhu[1*]
[1] Dept. of Comp. Sci. and Tech., BNRist Center, Institute for AI, Tsinghua-Bosch Joint ML Center
[1] Tsinghua University, Beijing, 100084, China [2] International Digital Economy Academy
{liusl20, yangxiao19}@mails.tsinghua.edu.cn, leizhang@idea.edu.cn, {suhangss, dcszj}@mail.tsinghua.edu.cn

2021

Background

Multi-label image classification is a visual recognition task that aims to predict a set of labels corresponding to objects, attributes, or actions given an input image.

Compared with single label classification, multi-label classification requires special attention :
1) how to handle the label imbalance problem.
2) how to extract features from region of interests.
3) How to find label correlations.



2 positive label: Person, Camera
78 negative labels
Missing label: Bottle

A typical image contains few positive samples, and many negative ones, leading to high negative-positive imbalance. Also, missing labels in ground-truth are common in multi-label datasets.
A novel asymmetric loss (ASL) enables to dynamically down-weights and hard-thresholds easy negative samples, while also discarding possibly mislabeled samples.

Ben-Baruch E, Ridnik T, Zamir N, et al. Asymmetric loss for multi-label classification[J]. arXiv preprint arXiv:2009.14119, 2020.
https://arxiv.org/pdf/2009.14119.pdf

Background

Asymmetric loss (ASL):

A general form of a binary loss per label,L, is given by:

$$L = -yL_+ - (1-y)L_-$$

ground-truth

Focal loss is obtained by setting $L^+$and L^-as:

$$\begin{cases} L_+ = (1-p)^{\gamma}\log(p) \\ L_- = p^{\gamma}\log(1-p) \end{cases}$$

The focusing parameter. $\gamma = 0$ yields binary cross-entropy

output probability

To define the Asymmetric Loss (ASL), we decouple the focusing levels of the positive and negative samples and integrate the two mechanisms of asymmetric focusing and probability shifting into a unified formula :

$$ASL = \begin{cases} L_+ = (1-p)^{\gamma_+}\log(p) \\ L_- = (p_m)^{\gamma_-}\log(1-p_m) \end{cases}$$

Adaptive Asymmetry, $\gamma_- \leftarrow \gamma_- + \lambda(\Delta p - \Delta p_{\text{target}})$

Asymmetric Probability Shifting, $p_m = \max(p-m, 0)$

Background

probability shifting performs hard thresholding of very easy negative samples, it fully discards negative samples when their probability is very low.

$$ASL = \begin{cases} L_+ = & (1-p)^{\gamma_+}\log(p) \\ L_- = & (p_m)^{\gamma_-}\log(1-p_m) \end{cases}$$

$$p_m = \max(p - m, 0)$$

$$\frac{dL_-}{dz} = \frac{\partial L_-}{\partial p}\frac{\partial p}{\partial z}$$
$$= (p_m)^{\gamma_-}\left[\frac{1}{1-p_m} - \frac{\gamma_-\log(1-p_m)}{p_m}\right]p(1-p)$$

Where $p = \frac{1}{1+e^{-z}}$,

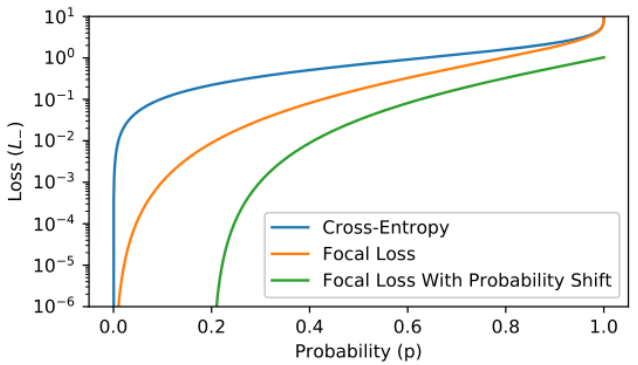


Figure 2: **Loss Comparisons.** Comparing probability-shifted focal loss to regular focal loss and cross-entropy, for negative samples. We used $\gamma_- = 2$ and $m = 0.2$.

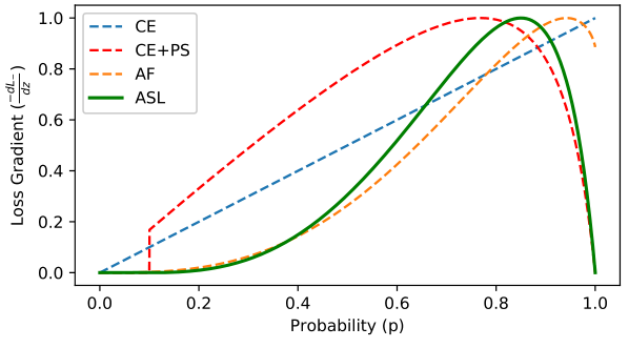


Figure 3: **Gradient Analysis.** Comparing the loss gradients vs. probability for different loss regimes. CE = Cross-Entropy ($m = \gamma_- = 0$), CE+PS = Cross-Entropy with Probability Shifting ($m > 0, \gamma_- = 0$), AF = Asymmetric Focusing ($m = 0, \gamma_- > 0$), ASL ($m > 0, \gamma_- > 0$).

Conclusion:

1.Hard-threshold - very easy negatives, with p < m, that should be ignored, in order to focus on harder samples.

2. Soft-threshold - negative samples, with p > m, that should be attenuated when their probability is low.

3. Mislabeled - very hard negative samples, with p > p*, where p* is defined as the point where $\frac{d}{dp}\left(\frac{dl}{dz}\right) = 0$,which are suspected as mislabeled

Background

Adaptive Asymmetry:

$$ASL = \begin{cases} L_+ = & (1-p)^{\gamma_+}\log(p) \\ L_- = & (p_m)^{\gamma_-}\log(1-p_m) \end{cases}$$

Adjust γ_- dynamically throughout the training, to match a desired probability gap, denoted by Δp_{target}.

$$\gamma_- \leftarrow \gamma_- + \lambda(\Delta p - \Delta p_{target}) \quad , \Delta p = p_t^+ - p_t^- .$$

p_t^+ and p_t^- : the average probabilities of the positive and negative samples, respectively

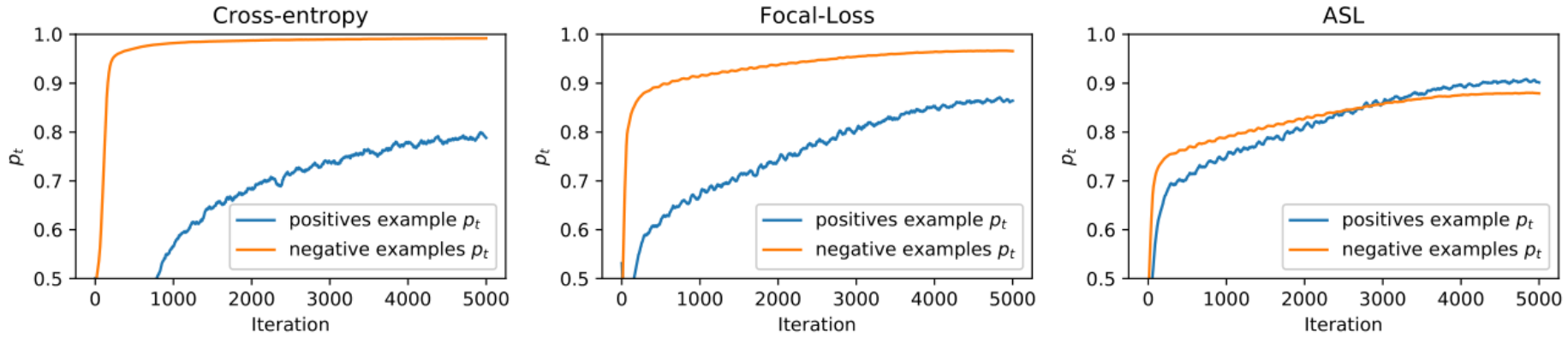


Figure 4: **Probability analysis.** The mean probability of positive and negative samples along the training with cross-entropy, focal loss and ASL, on MS-COCO. For focal loss we used $\gamma = 2$. For ASL we used $\gamma_+ = 0$, $\gamma_- = 2$, $m = 0.2$.

Conclusion:
enables us to dynamically increase the asymmetry level throughout the training, forcing the optimization process to focus more on the positive samples' gradients.

Background

Transformer:

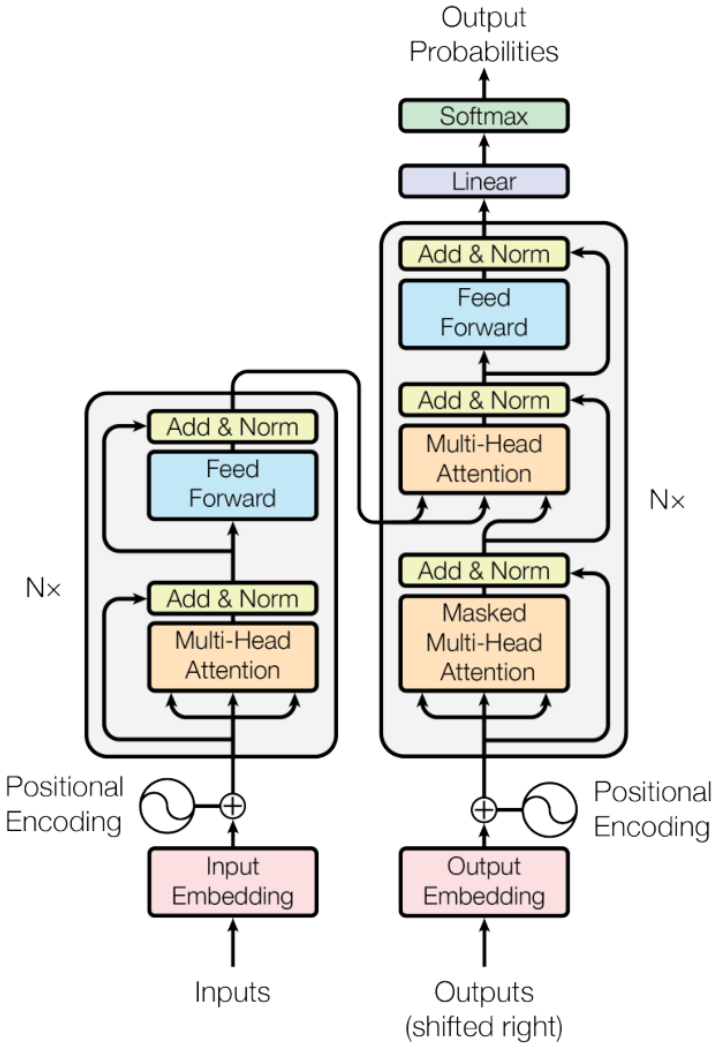
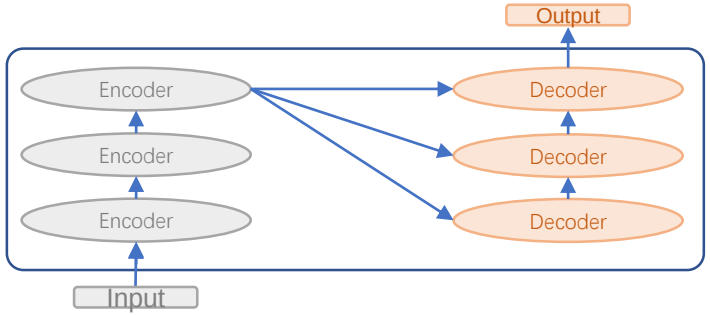


Figure 1: The Transformer - model architecture.



Encoder:

Each layer has two sub-layers. The first is a multi-head self attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection around each of the two sub-layers, followed by layer normalization.

Decoder:

a third sub-layer. We employ residual connections around each of the sub-layers, followed by layer normalization. We modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i.

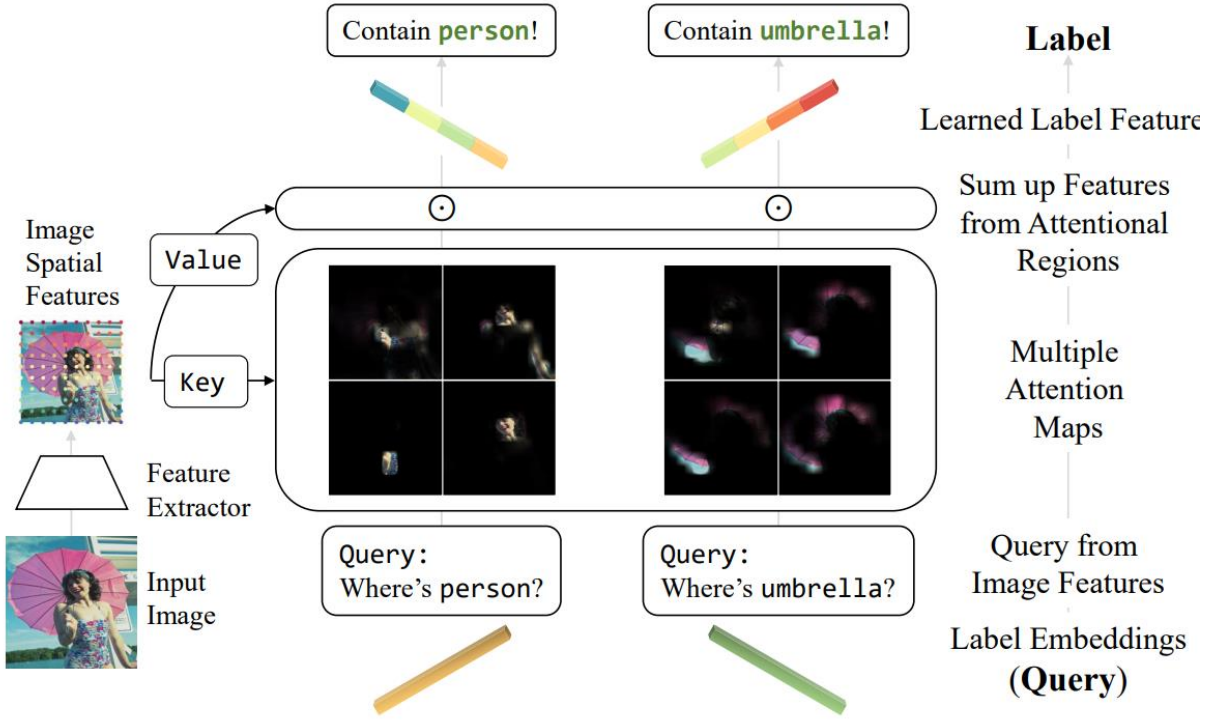
Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Query2Label:

As shown in the figure, we use learnable label embeddings as queries to probe and pool class-related features via the cross-attention module in Transformer encoders. The pooled features are adaptive and more discriminative, leading to a superior multi-label classification performance.



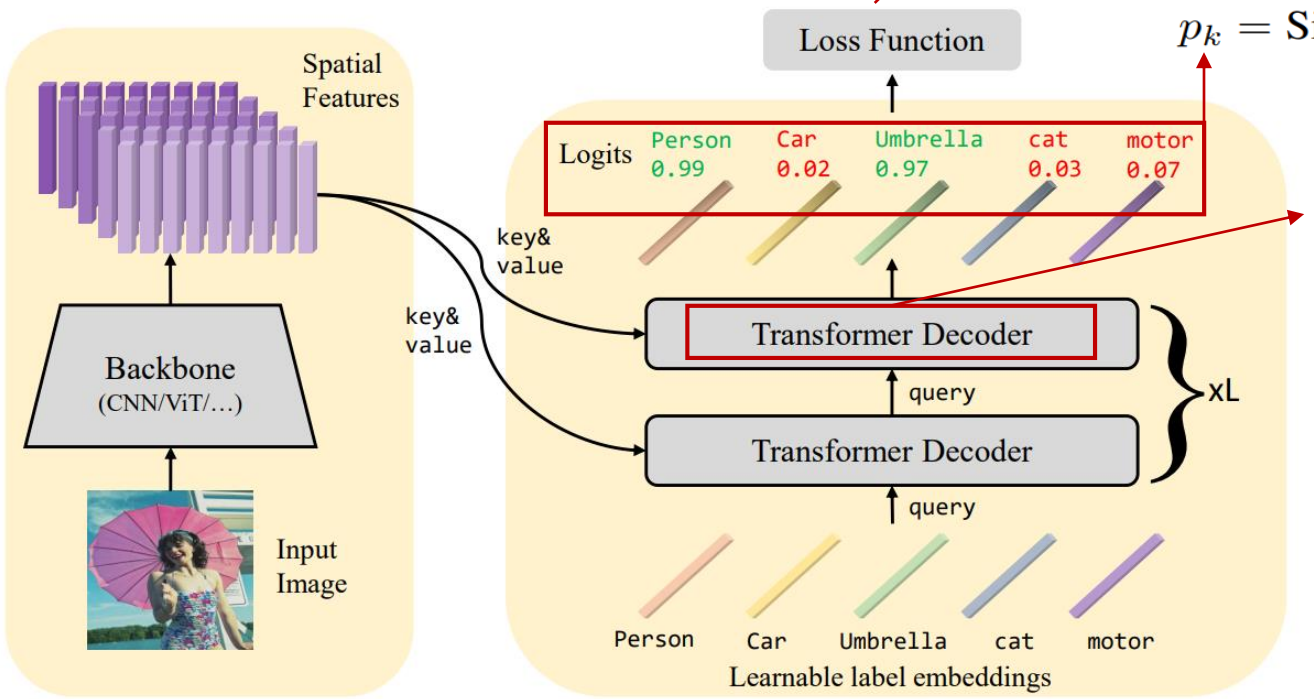
Using cross attention for adaptively feature pooling through focusing on different parts (best view in colors).

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$

Methods

For an input image, it firstly feeds it into a backbone in the first stage to extract spatial features.
The second stage is composed of two modules: a multi-layer Transformer decoder block for query updating and adaptive feature pooling, and a linear projection layer for computing prediction logits for each category.

$$ASL = \begin{cases} L_+ = & (1-p)^{\gamma_+}\log(p) \\ L_- = & (p_m)^{\gamma_-}\log(1-p_m) \end{cases}$$



$p_k = \text{Sigmoid}\left(W_k^T \mathcal{Q}_{L,k} + b_k\right)$,the predicted probabilities for each category

Query updating:

$$\text{self-attn:} \quad \mathcal{Q}_i^{(1)} = \text{MultiHead}(\widetilde{\mathcal{Q}}_{i-1}, \widetilde{\mathcal{Q}}_{i-1}, \mathcal{Q}_{i-1}),$$

$$\text{cross-attn:} \quad \mathcal{Q}_i^{(2)} = \text{MultiHead}(\widetilde{\mathcal{Q}}_i^{(1)}, \widetilde{\mathcal{F}}, \mathcal{F}),$$

$$\text{FFN:} \quad \mathcal{Q}_i = \text{FFN}(\mathcal{Q}_i^{(2)}),$$

~ means the original vectors modified by adding position encodings

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Ablation :

Method	small	medium	large
Baseline(TResNetL) [40]	37.8	74.2	84.2
Ours(TResNetL+Q2L)	39.5	77.5	86.1

Table 7: Comparison of improvement on objects with different sizes.

VOC2007 :

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [46]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
VGG+SVM [42]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
Fev+Lv [51]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
HCP [48]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RDAL [47]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RARL [6]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
SSGRL [7] (576)	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
MCAR [19]	99.7	**99.0**	98.5	98.2	85.4	96.9	97.4	98.9	83.7	95.5	88.8	99.1	98.2	95.1	99.1	84.8	97.1	87.8	98.3	94.8	94.8
ASL(TResNetL) [1]	**99.9**	98.4	98.9	98.7	86.8	98.2	98.7	98.5	83.1	**98.3**	**89.5**	98.8	**99.2**	98.6	**99.3**	89.5	**99.4**	86.8	**99.6**	95.2	95.8
ADD-GCN [52] (576)	99.8	**99.0**	98.4	**99.0**	86.7	98.1	98.5	98.3	**85.8**	**98.3**	88.9	98.8	99.0	97.4	99.2	88.3	98.7	**90.7**	99.5	**97.0**	96.0
Q2L-TResL(Ours)	**99.9**	98.9	**99.0**	98.4	**87.7**	**98.6**	**98.8**	**99.1**	84.5	**98.3**	89.2	**99.2**	**99.2**	**99.2**	**99.3**	**90.2**	98.8	88.3	99.5	95.5	**96.1**

Table 3: Comparisons of our method with previous state-of-the-art methods on PASCAL VOC 2007, in terms of AP and mAP in %. All results are reported at resolution 448×448 except for the ADD-GCN and SSGRL, whose resolutions are noted in parentheses. Results with advanced backbones could be found in the appendix.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

VOC2012 :

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGG+SVM [42]	99.0	89.1	96.0	94.1	74.1	92.2	85.3	97.9	79.9	92.0	83.7	97.5	96.5	94.7	97.1	63.7	93.6	75.2	97.4	87.8	89.3
Fev+Lv [51]	98.4	92.8	93.4	90.7	74.9	93.2	90.2	96.1	78.2	89.8	80.6	95.7	96.1	95.3	97.5	73.1	91.2	75.4	97.0	88.2	89.4
HCP [48]	99.1	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	94.7	71.9	96.7	88.6	90.5
MCAR [19]	99.6	97.1	98.3	96.6	87.0	95.5	94.4	98.8	87.0	96.9	85.0	98.7	98.3	97.3	99.0	83.8	96.8	83.7	98.3	93.5	94.3
SSGRL [7](576)	99.7	96.1	97.7	96.5	86.9	95.8	95.0	98.9	88.3	97.6	87.4	99.1	99.2	97.3	99.0	84.8	98.3	85.8	99.2	94.1	94.8
ADD-GCN [52](576)	99.8	97.1	98.6	96.8	89.4	97.1	96.5	99.3	89.0	97.7	87.5	99.2	99.1	97.7	99.1	86.3	**98.8**	87.0	99.3	95.4	95.5
Q2L-TResL(Ours)	**99.9**	**98.2**	**99.3**	**98.1**	**90.4**	**97.7**	**97.4**	**99.4**	**92.7**	**98.7**	**89.9**	**99.4**	**99.5**	**99.0**	**99.4**	**88.4**	**98.8**	**89.3**	**99.6**	**96.8**	**96.6**

NUS-WIDE:

Method	Backbone	mAP	CF1	OF1
MS-CMA [53]	ResNet101	61.4	60.5	73.8
SRN [56]	ResNet101	62.0	58.5	73.4
ICME [9]	ResNet101	62.8	60.7	74.1
Q2L-R101(Ours)	ResNet101	**65.0**	**63.1**	**75.0**
Baseline [40]	TresNetL	63.1	61.7	74.6
Focal loss [34]	TresNetL	64.0	62.9	74.7
ASL [1]	TresNetL	65.2	63.6	**75.0**
Q2L-TResL(Ours)	TresNetL	**66.3**	**64.0**	**75.0**
MlTr-l [10]	MlTr-l(22k)	66.3	65.0	75.8
Q2L-CvT(Ours)	CvT-w24(22k)	**70.1**	**67.6**	**76.3**

VG500:

Method	mAP
ResNet-101 [25]	30.9
ResNet-SRN [56]	33.5
SSGRL(ResNet101) [7]	36.6
C-Tran(ResNet101) [33]	38.4
Q2L-R101(Ours)	39.5
Q2L-TResL-22k(Ours)	**42.5**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

MS-COCO :

448×448(medium resolution):

Method	Backbone	Resolution	mAP	All						Top 3					
				CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
SRN [56]	ResNet101	224×224	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101 [25]	ResNet101	224×224	78.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
CADM [8]	ResNet101	448×448	82.3	82.5	72.2	77.0	84.0	75.6	79.6	87.1	63.6	73.5	89.4	66.0	76.0
ML-GCN [9]	ResNet101	448×448	83.0	85.1	72.0	78.0	85.8	75.4	80.3	87.2	64.6	74.2	89.1	66.7	76.3
KSSNet [36]	ResNet101	448×448	83.7	84.6	73.2	77.2	87.8	76.2	81.5	-	-	-	-	-	-
MS-CMA [53]	ResNet101	448×448	83.8	82.9	74.4	78.4	84.4	77.9	81.0	86.7	64.9	74.3	90.9	67.2	77.2
MCAR [20]	ResNet101	448×448	83.8	85.0	72.1	78.0	88.0	73.9	80.3	88.1	65.5	75.1	91.0	66.3	76.7
SSGRL [7]	ResNet101	576×576	83.8	89.9	68.5	76.8	91.3	70.8	79.7	91.9	62.5	72.7	93.8	64.1	76.2
C-Trans [33]	ResNet101	576×576	85.1	86.3	74.3	79.9	**87.7**	76.5	81.7	90.1	65.7	76.0	92.1	**71.4**	77.6
ADD-GCN [52]	ResNet101	576×576	85.2	84.7	75.9	80.1	84.9	**79.4**	82.0	88.8	66.2	75.8	90.3	68.5	77.9
Q2L-R101(Ours)	ResNet101	448×448	84.9	84.8	74.5	79.3	86.6	76.9	81.5	78.0	69.1	73.3	80.7	70.8	75.4
Q2L-R101(Ours)	ResNet101	576×576	**86.5**	**85.8**	**76.7**	**81.0**	87.0	78.9	**82.8**	**90.4**	**66.3**	**76.5**	**92.4**	67.9	**78.3**
ASL [1]	TResNetL	448×448	86.6	87.2	76.4	81.4	88.2	79.2	81.8	91.8	63.4	75.1	92.9	66.4	77.4
TResNetL [39]	TResNetL(22k)	448×448	88.4	-	-	-	-	-	-	-	-	-	-	-	-
Q2L-TResL(Ours)	TResNetL	448×448	87.3	**87.6**	76.5	81.6	**88.4**	78.5	83.1	**91.9**	66.2	77.0	**93.5**	67.6	78.5
Q2L-TResL(Ours)	TResNetL(22k)	448×448	**89.2**	86.3	**81.4**	**83.8**	86.5	**83.3**	**84.9**	91.6	**69.4**	**79.0**	92.9	**70.5**	**80.2**
MlTr-l [10]	MLTr-l(22k)	384×384	88.5	86.0	81.4	83.3	86.5	83.4	84.9	-	-	-	-	-	-
Swin-L [37]	Swin-L(22k)	384×384	89.6	**89.9**	80.2	84.8	**90.4**	82.1	86.1	93.6	69.9	80.0	94.3	71.1	81.1
CvT-w24 [49]	CvT-w24(22k)	384×384	90.5	89.4	81.7	85.4	89.6	83.8	86.6	93.3	70.5	80.3	94.1	71.5	81.3
Q2L-SwinL(Ours)	Swin-L(22k)	384×384	90.5	89.4	81.7	85.4	89.8	83.2	86.4	**93.9**	70.4	80.5	**94.8**	71.0	81.2
Q2L-CvT(Ours)	CvT-w24(22k)	384×384	**91.3**	88.8	**83.2**	**85.9**	89.2	**84.6**	**86.8**	92.8	**71.6**	**80.8**	93.9	**72.1**	**81.6**

Table 1: Comparison of our method with known state-of-the-art models on MS-COCO at medium input resolution. The backbones noted with 22k are pretrained on the ImageNet-22k dataset. Among them, mAP, OF1, and CF1 are the primary metrics (shaded in the table) as the others may be affected by the chosen threshold largely. All metrics are in %.

640×640(high resolution) :

Method	Architecture	Input Resolution	mAP
ASL [1]	TResNetXL	640×640	88.4
TResNet [39]	TResNetL(22k)	640×640	89.8
Q2L-TResXL	TResNetXL	640×640	89.0
Q2L-TResL	TResNetL(22k)	640×640	**90.3**

Table 2: Comparison of our method with ASL on MS-COCO for high input resolution of 640 × 640. All metrics are in %.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Visualization of Attention Maps:

Figure 4: Image examples classified correctly by Q2L but wrongly by the baseline TResNetL. The middle two columns are the mean attention maps of Q2L and the enlarged maps on focused regions respectively. The small scale of objects makes it difficult for TResNetL to recognize. Best view in colors.

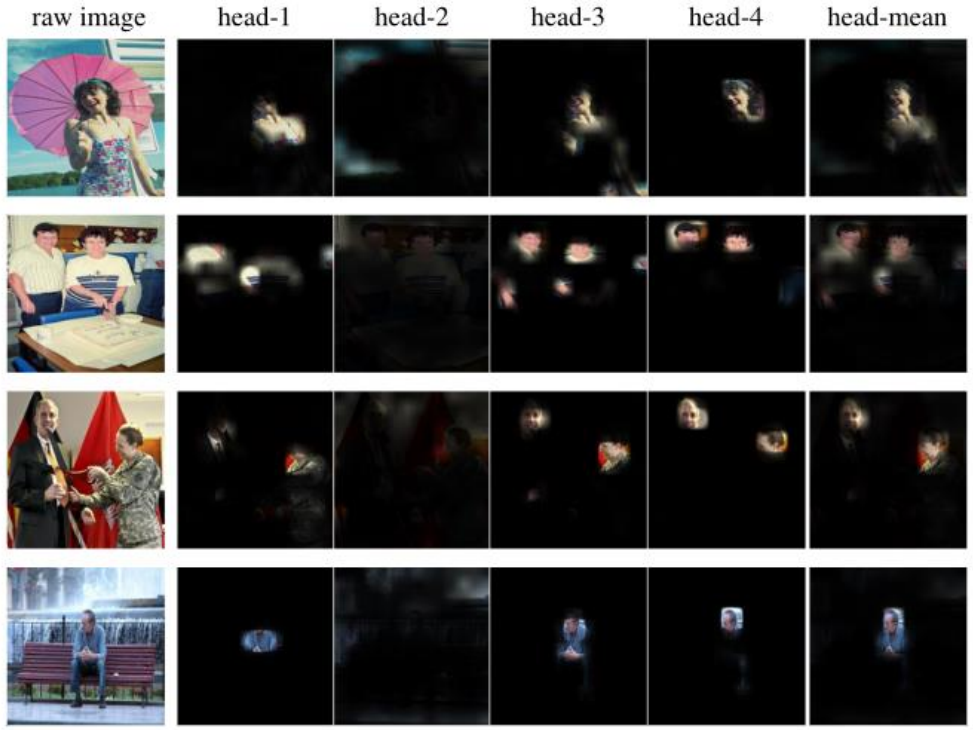


Figure 5: Visualization of multi-head attention maps for the target label `person`. Each column in the middle represents an attention map for one head and the rightmost column averages the maps of all heads. Best view in colors.

Thanks