CVPDL HW2

B09901142 EE4 呂睿超

November 2024

Generic Object Detection

1. How does DETR achieve end-to-end object detection using Transformers and bipartite matching? Briefly discuss the advantages and limitations of this approach compared to traditional convolutional detectors.

   (a) **Advantage:**

      i. It uses bipartite matching on object detection to transform the problem into a **direct set prediction** problem. Therefore, it eliminates the use of post-processing such as Non maximum suppression (NMS) to eliminate duplicates.

      ii. Also, utilizing Transformer architecture do global computations to do better on longer input sequences such as images.

   (b) **Limitations:**

      i. Compared to convolutional methods such as Fast-RCNN, it has **slower convergence** because of instability of bipartite matching and positional encoding issue.

      ii. Following the advantage of Transformer architecture on global computations, DETR performs better on large objects. However, it has slightly **lower performances on smaller objects**.

2. How does the DINO model improve training efficiency and accuracy over DETR? Briefly explain how "Contrastive DeNoising," "Mixed Query Selection," and "Look Forward Twice" enhance small object detection and overall AP on the COCO dataset.

   (a) **Contrastive DeNoising:**

      i. DeNoising techique feeds **noise-added ground truth** to train the model to reconstruct.

      ii. Contrastive DeNoising further improve the robustness by introducing **two queries**: positive and negative, enabling the model to learn to reject the negative ones, to **determine the empty boxes**.

      iii. The smaller objects are more likely to be disturbed by background noise and nearby object, so CDN mitigates the problem.

   (b) **Mixed Query Selection:**

      i. Mixed query selection is a **hybrid technique** that leave content queries static but initializes the anchor boxes using positional embedding of the top-K features from encoder.

    ii. Fixing the content queries is to **not mislead the decoder** by not refined content information. On the other hand, this hybrid model can use dynamic positional information to lead more comprehensive content information.

(c) **Look Forward twice:**

   i. In this technique, we refine the box prediction in the decoding process by *looking ahead*. The **i-th** layer box predictions would be determined by **(i-1)-th** layer **output box** and the **i-th** layer predicted box **offset**.

$$b_i^{(pred)} = Update(b'_{i-1}, \Delta b_i)$$

   ii. This improves the small object detection by enhancing the detail box prediction, which is more precise.

Practical Issues 1: Lightweight Computer Vision

3. How does the EdgeViT model achieve efficient performance on mobile devices compared to traditional Vision Transformers (ViTs) and CNNs, and what key architectural features contribute to this efficiency?

   (a) First, the EdgeViT is designed in a *hierarchical pyramid network structure*, which is to reduce spatial resolution but expand channels during each stage. This technique **sacrifices only a limited loss but improves the efficiency.**

   (b) EdgeViT introduces the ***Local-Global-Local bottleneck*** to largely reduce the attention computations required.

      i. First, the local aggregation collects the nearby tokens' information and aggregate to the center.

      ii. Next, the sparse attention **only do attention computation on those selected *center tokens.*** This achieves a quadratic computation reduce.

      iii. Lastly, the local propagation utilizes the transposed convolution to effectively *spread* the information back to all the tokens.

      iv. Therefore, EdgeViT successfully reduces the computation needed and reserves the global information that Transformers have.

4. What is the main design strategy of the Lite DETR model to improve efficiency in Transformer-based object detection, and how does it achieve computational savings while maintaining performance?

   (a) The main strategy that Lite DETR introduced, is to utilize a **interleaving update of high and low level features**

      i. First, we obtain different level features S_1 to S_4 from different layers of the backbone(e.g. ResNet-50).

      ii. Next, we only do query computation of the h**igh-level features for most times** in the encoder block. On the other hand, we only do one time of the query computation of the low-level features.

      iii. This strategy achieve computational savings **because the high-level features are only a small (25%) portion.** Meanwhile, the last time of the computation ensures that the low-level features are also updated.

Practical Issue 2: Data Imbalance and Domain Adaptation

5. Based on the concepts discussed in “Class-Balanced Loss Based on Effective Number of Samples” explain the ”effective number of samples” concept and how it helps address challenges associated with long-tailed data distributions.

   (a) First, this paper introduces the **Effective number of samples**, which determines the truly useful information of the data. It is derived from the probability of newly sampled data to overlap previous ones.

   (b) Next, the paper integrates the **effective number** into the re-weighting technique by directly dividing the weighting. The formula is as below (e.g. loss)

   $$CB(p, y) = \frac{1}{E_{n_y}} L(p, y)$$

   Note that minority classes have larger E_{n_y}.

   (c) Thus, the long-tailed data distributions can be mitigated it truly determines the effective samples and c**orrectly modified the loss function by the re-weighting.**

6. How does Adversarial Domain Alignment improve model performance in domain adaptation tasks?

   (a) Whether feature-space or pixel-space, Adversarial Domain Alignment aims to **align the source and target distributions**. If the two distributions are same, then the classification (boundary) trained on source can be directly used on target data.

   (b) For feature-space method, the paper introduces the **domain discriminator** to identify which domain does a encoded data come from. As we finished training the framework, the target encoder should output **embeddings that the discriminator could not figure out which domain** does it come from.

   (c) For pixel-space method, we train a converter to convert source data to align with target data also with a discriminator to determine whether the convert is effective.

Practical Issue 3: Weakly-/ Semi-supervised Learning

7. What is the primary difference between weakly supervised and semi-supervised learning in terms of annotation requirements for training in computer vision tasks?

   (a) Weakly supervised learning is to annotate the data with *weak label.* Weak label refers to label that does not reach the task requirement. For instance, if we want to do object detection but the data is only labeled with the object classes without labeling the bounding box.

   (b) Semi-supervised learning is to annotate only a portion of the data but with label that meets requirement. The rest of the data are remained unannotated.

8. In the context of Multiple-Instance Learning (MIL) applied to weakly supervised object detection (WSOD), what is the purpose of using positive and negative "bags," and what challenges does this approach present?

    (a) **Positive bags** denote the images with the target object, while **negative bags** does not contain the target object. Because weakly supervised annotation only contain class label, thus using positive and negative bags and compare the patches in those bags could create a better method.

    (b) Typical MIL strategy starts from initializing from the patches of positive bags and go into loops of re-training and re-localization

    (c) However, there are some challenge of the approach.

        i. First, the **initialization** is a hard problem because we don't know the detail relationship between patches across positive bags

        ii. Second, the **optimal boundary** of bag related method isn't the optimal of instance based method.

9. In the MixMatch algorithm for semi-supervised learning, how does the use of data augmentation and label guessing contribute to reducing the reliance on labeled data, and why are these techniques effective in leveraging unlabeled data?

    (a) **Data augmentation:**

        i. Adopting data augmentation makes effective the labeled data larger

        ii. Multiple augmentations of the unlabeled data enhances the model to **gain the knowledge of some variance** of the data.

        iii. Mixing all these labeled and unlabeled data with augmentations together create a robust understanding toward data without label.

    (b) **Label Guessing:**

        i. First, adopting the average of all the predictions create a confident prediction among augmentations.

        ii. The improvement of label guessing that this paper introduces is the use of a post-processing : **Temperature Scaling**. Temperature scaling is a technique derived from the concept of entropy minimization. This **sharpens the prediction distribution** and creates a more **confident and reliable guess.**

Practical Issue 4: Self-supervised Learning

10. SimCLR's contrastive learning framework achieves high performance by leveraging a particular loss function known as NT-Xent. Explain the mechanism behind the NT-Xent loss and discuss why it is particularly effective for self-supervised learning.

    (a) NT-Xent is a loss design to let similar **images to attract and different images to rebel.** More detailed implementation are as following

        i. SimCLR creates one augmentation of each image, creating a pair. Next, the pair goes through the encoder and the MLP to form a representation embedding z_i

ii. Then by calculating similarity of z_i, z_j and use them in the design of NT-Xent, we can achieve the objective to let the encoder create embeddings that attract images that are from the same pair.

iii. This design doesn't require label since it depends on similarity calculation.

(b) Also, the similarity design is modified by **tuning a temperature hyper-parameter** τ, so that the similarity is reasonable and effective.

(c) Therefore, this design no only doesn't require label but also utilizes the large pool dissimilar images, which is particularly effective for self-supervised learning and for downstream task.