Homework 4

Instructions

- This homework focuses on understanding and applying CoCoOp for CLIP prompt tuning. It consists of **four questions** designed to assess both theoretical understanding and practical application.
- Please organize your answers and results for the questions below and submit this jupyter notebook as **a .pdf file**.
- **Deadline: 11/26 (Tue) 23:59**

Preparation

- Run the code below before proceeding with the homework (Q1, Q2).
- If an error occurs, click 'Run Session Again' and then restart the runtime from the beginning.

[ ] ↳ *숨겨진 셀 1개*

Q1. Understanding and implementing CoCoOp

- We have learned how to define CoOp in Lab Session 4.
- The main difference between CoOp and CoCoOp is **meta network** to extract image tokens that is added to the text prompt.
- Based on the CoOp code given in Lab Session 4, fill-in-the-blank exercise to test your understanding of critical parts of the CoCoOp.

```
import torch.nn as nn

class CoCoOpPromptLearner(nn.Module):
    def __init__(self, cfg, classnames, clip_model):
        super().__init__()
        n_cls = len(classnames)
        n_ctx = cfg.TRAINER.COCOOP.N_CTX
        ctx_init = cfg.TRAINER.COCOOP.CTX_INIT
        dtype = clip_model.dtype
        ctx_dim = clip_model.ln_final.weight.shape[0]
        vis_dim = clip_model.visual.output_dim
        clip_imsize = clip_model.visual.input_resolution
        cfg_imsize = cfg.INPUT.SIZE[0]
        assert cfg_imsize == clip_imsize, f"cfg_imsize ({cfg_imsize}) must equal to clip_imsize ({clip_imsize})"

        if ctx_init:
            # use given words to initialize context vectors
            ctx_init = ctx_init.replace("_", " ")
            n_ctx = len(ctx_init.split(" "))
            prompt = clip.tokenize(ctx_init)
            with torch.no_grad():
                embedding = clip_model.token_embedding(prompt).type(dtype)
            ctx_vectors = embedding[0, 1: 1 + n_ctx, :]
            prompt_prefix = ctx_init
        else:
            # random initialization
            ctx_vectors = torch.empty(n_ctx, ctx_dim, dtype=dtype)
            nn.init.normal_(ctx_vectors, std=0.02)
            prompt_prefix = " ".join(["X"] * n_ctx)

        print(f'Initial context: "{prompt_prefix}"')
        print(f"Number of context words (tokens): {n_ctx}")

        self.ctx = nn.Parameter(ctx_vectors)  # Wrap the initialized prompts above as parameters to make them trainable.

        ### Tokenize ###
        classnames = [name.replace("_", " ") for name in classnames]  # 예) "Forest"
        name_lens = [len(_tokenizer.encode(name)) for name in classnames]
        prompts = [prompt_prefix + " " + name + "." for name in classnames] # 예) "A photo of Forest."

        tokenized_prompts = torch.cat([clip.tokenize(p) for p in prompts]) # 예) [49406, 320, 1125, 539...]



        ####################################
        ####### Q1. Fill in the blank #######
        ########## Define Meta Net ##########
        self.meta_net = nn.Sequential(OrderedDict([
            ("linear1", nn.Linear(vis_dim, vis_dim//16)),
            ("relu", nn.ReLU(inplace=True)),
```

```
            ("linear2", nn.Linear(vis_dim // 16, ctx_dim))
        ]))
        ####################################
        ## Hint: meta network is composed to linear layer, relu activation, and linear layer.



        if cfg.TRAINER.COCOOP.PREC == "fp16":
            self.meta_net.half()

        with torch.no_grad():
            embedding = clip_model.token_embedding(tokenized_prompts).type(dtype)

        # These token vectors will be saved when in save_model(),
        # but they should be ignored in load_model() as we want to use
        # those computed using the current class names
        self.register_buffer("token_prefix", embedding[:, :1, :])  # SOS
        self.register_buffer("token_suffix", embedding[:, 1 + n_ctx:, :])  # CLS, EOS
        self.n_cls = n_cls
        self.n_ctx = n_ctx
        self.tokenized_prompts = tokenized_prompts  # torch.Tensor
        self.name_lens = name_lens

    def construct_prompts(self, ctx, prefix, suffix, label=None):
        # dim0 is either batch_size (during training) or n_cls (during testing)
        # ctx: context tokens, with shape of (dim0, n_ctx, ctx_dim)
        # prefix: the sos token, with shape of (n_cls, 1, ctx_dim)
        # suffix: remaining tokens, with shape of (n_cls, *, ctx_dim)

        if label is not None:
            prefix = prefix[label]
            suffix = suffix[label]

        prompts = torch.cat(
            [
                prefix,  # (dim0, 1, dim)
                ctx,  # (dim0, n_ctx, dim)
                suffix,  # (dim0, *, dim)
            ],
            dim=1,
        )

        return prompts

    def forward(self, im_features):
        prefix = self.token_prefix
        suffix = self.token_suffix
        ctx = self.ctx  # (n_ctx, ctx_dim)



        ############################################
        ########## Q2,3. Fill in the blank #########
        bias=self.meta_net(im_features)  # (batch, ctx_dim)
        bias = bias.unsqueeze(1)  # (batch, 1, ctx_dim)
        ctx = ctx.unsqueeze(0)  # (1, n_ctx, ctx_dim)
        ctx_shifted=ctx+bias  # (batch, n_ctx, ctx_dim)
        ############################################
        ############################################



        # Use instance-conditioned context tokens for all classes
        prompts = []
        for ctx_shifted_i in ctx_shifted:
            ctx_i = ctx_shifted_i.unsqueeze(0).expand(self.n_cls, -1, -1)
            pts_i = self.construct_prompts(ctx_i, prefix, suffix)  # (n_cls, n_tkn, ctx_dim)
            prompts.append(pts_i)
        prompts = torch.stack(prompts)

        return prompts
```

```
class CoCoOpCustomCLIP(nn.Module):
    def __init__(self, cfg, classnames, clip_model):
        super().__init__()
        self.prompt_learner = CoCoOpPromptLearner(cfg, classnames, clip_model)
        self.tokenized_prompts = self.prompt_learner.tokenized_prompts
        self.image_encoder = clip_model.visual
        self.text_encoder = TextEncoder(clip_model)
        self.logit_scale = clip_model.logit_scale
        self.dtype = clip_model.dtype

```

```
    def forward(self, image, label=None):
        tokenized_prompts = self.tokenized_prompts
        logit_scale = self.logit_scale.exp()

        image_features = self.image_encoder(image.type(self.dtype))
        image_features = image_features / image_features.norm(dim=-1, keepdim=True)


        ############################################
        ########## Q4. Fill in the blank #########
        prompts=self.prompt_learner(image_features)
        ############################################
        ############################################


        logits = []
        for pts_i, imf_i in zip(prompts, image_features):
            text_features = self.text_encoder(pts_i, tokenized_prompts)
            text_features = text_features / text_features.norm(dim=-1, keepdim=True)
            l_i = logit_scale * imf_i @ text_features.t()
            logits.append(l_i)
        logits = torch.stack(logits)

        if self.prompt_learner.training:
            return F.cross_entropy(logits, label)

        return logits
```

Q2. Trainining CoCoOp

In this task, you will train CoCoOp on the EuroSAT dataset. If your implementation of CoCoOp in Question 1 is correct, the following code should execute without errors. Please submit the execution file so we can evaluate whether your code runs without any issues.

```
# Train on the Base Classes Train split and evaluate accuracy on the Base Classes Test split.
args.trainer = "CoCoOp"
args.train_batch_size = 4
args.epoch = 100
args.output_dir = "outputs/cocoop"

args.subsample_classes = "base"
args.eval_only = False
cocoop_base_acc = main(args)
```

```
Loading trainer: CoCoOp
Loading dataset: EuroSAT
Reading split from /content/ProMetaR/ProMetaR/data/eurosat/split_zhou_EuroSAT.json
Creating a 16-shot dataset
Creating a 4-shot dataset
Saving preprocessed few-shot data to /content/ProMetaR/ProMetaR/data/eurosat/split_fewshot/shot_16-seed_1.pkl
SUBSAMPLE BASE CLASSES!
Building transform_train
+ random resized crop (size=(224, 224), scale=(0.08, 1.0))
+ random flip
+ to torch tensor of range [0, 1]
+ normalization (mean=[0.48145466, 0.4578275, 0.40821073], std=[0.26862954, 0.26130258, 0.27577711])
Building transform_test
+ resize the smaller edge to 224
+ 224x224 center crop
+ to torch tensor of range [0, 1]
+ normalization (mean=[0.48145466, 0.4578275, 0.40821073], std=[0.26862954, 0.26130258, 0.27577711])
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:617: UserWarning: This DataLoader will create 8 worker processes in to
  warnings.warn(
---------  -------
Dataset    EuroSAT
# classes  5
# train_x  80
# val      20
# test     4,200
---------  -------
Loading CLIP (backbone: ViT-B/16)
Building custom CLIP
Initial context: "a photo of a"
Number of context words (tokens): 4
Turning off gradients in both the image and the text encoder
Parameters to be updated: {'prompt_learner.meta_net.linear2.weight', 'prompt_learner.meta_net.linear2.bias', 'prompt_learner.meta_net.linear1
Loading evaluator: Classification
No checkpoint found, train from scratch
/usr/local/lib/python3.10/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning: The verbose parameter is deprecated. Please use get_last_
  warnings.warn(
epoch [1/100] batch [20/20] time 0.100 (0.333) data 0.000 (0.047) loss 0.2744 (1.1881) lr 2.5000e-03 eta 0:10:58
epoch [2/100] batch [20/20] time 0.091 (0.123) data 0.000 (0.017) loss 0.8384 (0.8970) lr 2.4994e-03 eta 0:04:00
epoch [3/100] batch [20/20] time 0.092 (0.122) data 0.000 (0.020) loss 0.6382 (0.7859) lr 2.4975e-03 eta 0:03:55
epoch [4/100] batch [20/20] time 0.092 (0.120) data 0.000 (0.023) loss 0.5044 (0.7151) lr 2.4945e-03 eta 0:03:50
```

```
epoch [5/100] batch [20/20] time 0.136 (0.188) data 0.000 (0.034) loss 0.5703 (0.6317) lr 2.4901e-03 eta 0:05:56
epoch [6/100] batch [20/20] time 0.090 (0.124) data 0.000 (0.025) loss 0.6060 (0.6009) lr 2.4846e-03 eta 0:03:52
epoch [7/100] batch [20/20] time 0.090 (0.141) data 0.000 (0.016) loss 0.3853 (0.6638) lr 2.4779e-03 eta 0:04:21
epoch [8/100] batch [20/20] time 0.092 (0.123) data 0.000 (0.017) loss 1.4082 (0.6633) lr 2.4699e-03 eta 0:03:46
epoch [9/100] batch [20/20] time 0.120 (0.137) data 0.000 (0.019) loss 0.1780 (0.4582) lr 2.4607e-03 eta 0:04:09
epoch [10/100] batch [20/20] time 0.136 (0.209) data 0.000 (0.033) loss 1.2285 (0.5051) lr 2.4504e-03 eta 0:06:16
epoch [11/100] batch [20/20] time 0.094 (0.130) data 0.000 (0.017) loss 0.2539 (0.5013) lr 2.4388e-03 eta 0:03:50
epoch [12/100] batch [20/20] time 0.091 (0.124) data 0.000 (0.021) loss 1.1484 (0.4657) lr 2.4261e-03 eta 0:03:38
epoch [13/100] batch [20/20] time 0.100 (0.122) data 0.000 (0.015) loss 0.8467 (0.5009) lr 2.4122e-03 eta 0:03:32
epoch [14/100] batch [20/20] time 0.129 (0.137) data 0.000 (0.019) loss 0.5547 (0.4495) lr 2.3972e-03 eta 0:03:56
epoch [15/100] batch [20/20] time 0.134 (0.189) data 0.000 (0.032) loss 1.0430 (0.5549) lr 2.3810e-03 eta 0:05:21
epoch [16/100] batch [20/20] time 0.092 (0.123) data 0.000 (0.018) loss 1.3906 (0.4799) lr 2.3638e-03 eta 0:03:26
epoch [17/100] batch [20/20] time 0.090 (0.122) data 0.000 (0.019) loss 0.0238 (0.3497) lr 2.3454e-03 eta 0:03:22
epoch [18/100] batch [20/20] time 0.095 (0.121) data 0.000 (0.019) loss 0.1337 (0.2804) lr 2.3259e-03 eta 0:03:18
epoch [19/100] batch [20/20] time 0.134 (0.133) data 0.000 (0.017) loss 1.0420 (0.3864) lr 2.3054e-03 eta 0:03:35
epoch [20/100] batch [20/20] time 0.139 (0.184) data 0.000 (0.025) loss 0.3484 (0.4984) lr 2.2839e-03 eta 0:04:54
```

```
# Accuracy on the New Classes.
args.model_dir = "outputs/cocoop"
args.output_dir = "outputs/cocoop/new_classes"
args.subsample_classes = "new"
args.load_epoch = 100
args.eval_only = True
cocoop_novel_acc = main(args)
```

```
Loading trainer: CoCoOp
Loading dataset: EuroSAT
Reading split from /content/ProMetaR/ProMetaR/data/eurosat/split_zhou_EuroSAT.json
Loading preprocessed few-shot data from /content/ProMetaR/ProMetaR/data/eurosat/split_fewshot/shot_16-seed_1.pkl
SUBSAMPLE NEW CLASSES!
Building transform_train
+ random resized crop (size=(224, 224), scale=(0.08, 1.0))
+ random flip
+ to torch tensor of range [0, 1]
+ normalization (mean=[0.48145466, 0.4578275, 0.40821073], std=[0.26862954, 0.26130258, 0.27577711])
Building transform_test
+ resize the smaller edge to 224
+ 224x224 center crop
+ to torch tensor of range [0, 1]
+ normalization (mean=[0.48145466, 0.4578275, 0.40821073], std=[0.26862954, 0.26130258, 0.27577711])
---------  -------
Dataset    EuroSAT
# classes  5
# train_x  80
# val      20
# test     3,900
---------  -------
Loading CLIP (backbone: ViT-B/16)
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:617: UserWarning: This DataLoader will create 8 worker processes in total
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning: The verbose parameter is deprecated. Please use get_last_lr
  warnings.warn(
/content/ProMetaR/dassl/utils/torchtools.py:102: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value)
  checkpoint = torch.load(fpath, map_location=map_location)
Building custom CLIP
Initial context: "a photo of a"
Number of context words (tokens): 4
Turning off gradients in both the image and the text encoder
Parameters to be updated: {'prompt_learner.meta_net.linear2.bias', 'prompt_learner.meta_net.linear2.weight', 'prompt_learner.meta_net.linear1.we
Loading evaluator: Classification
Loading weights to prompt_learner from "outputs/cocoop/prompt_learner/model.pth.tar-100" (epoch = 100)
Evaluate on the *test* set
100%|██████████| 39/39 [01:04<00:00,  1.66s/it]=> result
* total: 3,900
* correct: 1,687
* accuracy: 43.3%
* error: 56.7%
* macro_f1: 39.0%
```

Q3. Analyzing the results of CoCoOp

Compare the results of CoCoOp with those of CoOp that we trained in Lab Session 4. Discuss possible reasons for the performance differences observed between CoCoOp and CoOp.

CoOp and CoCoOp both investigates the most efficient input phrases instead of fixing format like 'a photo of {object}'.

However, while CoOp focuses solely on phrases and classes, CoCoOp utilises information from the image file. It prevents overfitting problem in CoOp, and this is implemented at the meta network; the homework part of Q1.

The differences can result in following consequences:

- CoOp gets higher accuracy when faced on base classes.
- CoCoOp gets higher accuracy when faced on unseen classes than CoOp.
- In overall situation, CoCoOp shows better performance.

Comparing the above with the result of CoOp in the Lab session 4, it performs worse in all situations. (90.8% < 92.3%, 43.3% < 51.5%)

Nevertheless, it does not mean that it is a counterexample against the claim. In this project, it sets learnable context as 4 words, especially regarding 'a photo of a' as a default state. On the other hand, CoOp implemented in Lab session 4 suggests learnable context as 16 words and it starts from none-state (blank).

Meanwhile, the speed of CoOp is faster than CoCoOp. This can be seen in runtime comparison.