1. Problem Definition

빠르게 변화하는 머신러닝 분야에서, 다양한 종류의 태스크들을 효율적으로 병렬 처리할 수 있는 모델의 필요성은 점차 증가해 왔다. 전통적인 머신 러닝의 접근법은 싱글 태스크 러닝으로 모델이 하나의 태스크만을 한번에 처리하는 것이다. 그러나 이는 다면적이고 상호 연관적인 태스크들이 혼재하는 현실 세계에서 그 효율성과 적용가능성으로 스스로 제약하는 측면이 있다. 이러한 문제는 **멀티 태스크 모델의** 발전을 꾀하였다. 멀티 태스크 모델은 다양한 태스크들의 공통된 모달리티와 차이점들을 이해하고 이를 효과적으로 학습 및 인퍼런스 할 수 있는 장점이 존재한다.

**프롬프트 튜닝이란** 다른 도메인 데이터에서 이미 학습된 언어 모델을 다운 스트림 태스크에 효과적으로 적용할 수 있도록 고안된 파라미터 튜닝이다. 이는 전체 파라미터를 튜닝하지 않고 효과적으로 일부 파라미터만 튜닝할 수 있게 함으로써 다운스트림 태스크의 학습을 효율적으로 할 수 있게 하였다. 그러나 프롬프트 튜닝은 자연어 처리 분야의 트랜스포머 기반 대형 언어 모델의 발전에서 탄생했기에 비전 태스크에의 적용은 한계점이 존재하였다. 그러나 VIT[1] 와 이를 기반으로 한VPT[2] 의 발전으로 비전 모델에서도 프롬프트 튜닝을 효과적으로 적용할 수 있게 되었다. 풍부한 데이터를 소수의 파라미터만으로 모델에 튜닝할 수 있는 프롬프트 튜닝 기법은 싱글 태스크 모델의 경우에는 효과적이나 모델이 멀티 태스크인 경우에는 서로 다른 태스크의 성질을 하나의 프롬프트에 효과적으로 반영하기 어렵다. 이에 따라 여러 프롬프트를 사용하여 태스크에 맞는 프롬프트를 사용하거나, 각 태스크 간 공통된 성질을 반영하는 프롬프트 튜닝 기법이 소개되었다 [3,4,5].

그러나, 해당 방법들은 자연어처리 분야에서 적용되거나 하나의 태스크 분야에서 다양한 다른 데이터 셋을 활용한 것이기에 완전한 의미의 멀티 태스크라 보기엔 어렵다. 따라서 본 연구는 비전 분야에서 서로 다른 성질을 갖는 태스크인 **Depth Estimation**과 **Segmentation**을 대상으로 하였다. 또한 서로 다른 태스크가 주어질 때 프롬프트 풀에서 해당 프롬프트를 가져오는 **Continual learning**방식[6]에서 영감을 얻어 서로 다른 태스크의 공통된 성질을 프롬프트에 담아 효과적인 학습을 유도하였다 . 현재 비전 분야에서 멀티 태스크의 중요성이 나날이 증가함에도 관련된 프롬프트 튜닝 기법이 충분히 연구되지 않았기에 본 연구의 **멀티 태스크와 프롬프트 튜닝의 통합은** 상당한 의미가 있다. 이러한 접근법은 기존의 전통적인 싱글 태스크 모델의 한계를 극복하고 멀티 태스크 모델의 적용 가능성을 다양한 스펙트럼의 세계로 확장할 것이다.

1. Data

본 연구는 segmentaion과 depth estimation 분야에서 널리 활용되는 벤치마크인 NYU Depth Dataset V2 (NYUv2)를 활용하였다. 해당 데이터 셋은 다양한 내부 환경에서 수집된 RGB와 Depth, Segmentation 이미지로 구성되어 있다. NYUv2 데이터 셋은 학습과 검증 단계에서segmentation 과 depth estimation 태스크의 풍부한 데이터 소스를 제공하므로 본 연구가 진행하는 Depth estimation 과 segmentation멀티 태스크 러닝에 특히 효과적으로 적용될 수 있다. NYUv2를 멀티 태스크 러닝에 적용한 연구는 다음과 같다 [7,8]. 이들 연구는 어떻게 멀티 태스크 모델들이NYUv2 데이터 셋을 통해 효과적으로 훈련되고 평가될 수 있는 지를 선보였다.

1. **Methods**

3.3. Depth estimation

본 연구는 DPT 헤드 [9] 를 decode layer에 활용했다. 학습 시엔 reverse Huber loss[10]를 활용했다.

Evaluation 시엔 테스트 데이터 셋에 대해 δ1 metric를 활용하여 1.25보다 낮은 error max{ yˆp/ yp , yˆp /yp} 범위 안에 있는 픽셀 p의 비율을 검증했다. 이외에도 추가적으로 δ2 (1.25\*\*2) , δ3(1.25\*\*3) 값을 측정하여 검증하였다.

1. Experiments

본 연구 실험은 Multi Masked Auto Encoder(MMAE) model 을 대상으로 한다. 해당 모델은 Masked auto encoder 모델을 다양한 모달리티와 태스크에서 활용하기 위해 고안된 모델이다 [7]. 해당 모델으 ImageNet dataset 1000k 에서 pretrain 되었다. 본 실험은 크게 5단계로 구성되어 있으며 각각은 depth estimation과 semantic segmentation 태스크의 다양한 환경 속에서 모델의 성능 평가로 목표로 진행되었다.

* 1. Single Task: Fine tuning MMAE for Depth Estimation.

우선, MMAE 의 성능을 검증하기 위해 single modality 환경에서 rgb 이미지 값을 인풋으로 한 single task finetuning을 진행하였다. 다운스트림 태스크는 depth estimation으로 진행하였다. 모델의 성능은 Delta 1 metric으로 평가되었다. 본 연구의 fine-tuned model 의 delta 1 값은 약 86.4 로 기존 논문의 보고된 성능과 거의 일치한다. 이 단계는 depth estimation 의 모델의 수행능력의 베이스 라인을 검증하는 것을 목표로 수행되었다.

* 1. Single Task: Fine tuning MMAE for Segmentation.
  2. Multi Task: Fine tuning MMAE for Multi-Task learning

그 후, 세번째 단계에서는 Depth estimation 과 segmentation 을 통합해서 한 에포크 당 각 태스크를 동시에 훈련할 수 있는 모델을 설계하였다. 이를 위해 pretrained 된 인코더에 각 태스크에 맞는 디코더를 붙여서 모델을 구성하였다. Segmentation에는 convNeXt 헤드를 사용하였고 depth esimation에는 DPT head를 붙여 모델을 구성하였다. 이어서 해당 MMAE 모델은 오직 RGB 값만을 인풋 모달리티로 활용해 파인 튜닝 되었다. 그러나 아웃풋 태스크는 depth 와 segmentation모두를 활용하였다. Semantic segmentation 에서의 성능은 Mean Intersection Over Union(mIoU) metric 을 활용해 평가되었고 depth esimation은 delta 1 metric을 통해 검증되었다. 모델은 mIoU 값으로 약 52를 달성하였다. 이 수치는 싱글 모달리티 싱글 태스크로 파인 튜닝 시킨 기존 논문의 성능과 일치한다. 해당 단계는 멀티 태스크를 수행하더라도 성능 저하가 일어나지 않음을 검증하는 것을 목표로 수행되었다.

* 1. Implementing Deep Prompt in MMAE

이 단계에서는 multi task 모델에서prompt tuning 방식의 효율성을 검증하고prompt pool 방식의 당위성을 검증하기 위해 deep prompt tuning이 수행되었다. 태스크는depth estimation 과 semantic segmentation을 대상으로 진행되었다. mIoU 값과 Delta 1값은 각각 33, 82가 나왔으며 이는 기존 full finetuning 시Delta 1값과 근접한 수치이다. 그러나 Segmentation 의 경우 33이 나왔는 데 이는 각 depth 와 segmentation loss의 normalization을 적절히 수행하지 못한 것이 원인으로 파악된다. 추후의 prompt pool 방식은 normalization을 진행하였다. 해당 결과는 multi task 모델에서의 prompt tuning의 잠재된 활용 가능성 및 을 효율성을 입증한다. 총 parameter 수는 full finetuning 시에 비해 30M로 약 1/3이상이 줄어들었다. Prompt 는 1개를 각 테스크가 공유하도록 하였고 length를 100으로 설정하여 학습시켰다. Single GPU A100 기준 200 에포크를 돌렸으며 총 학습시간은 약 12시간이 소요됐다.

* 1. Implementing Prompt Pool in MMAE

1. Result

|  |  |  |  |
| --- | --- | --- | --- |
| Deep visual prompt(length:100, No normalization) | 30M | 82.98 | 33 |

1. Discussion for Future Research
   1. Task specific prompt pooling

현재 구현된 prompt pooling 방식은 Training 및 Inference 시에 각 태스크의 헤드들이 같은 prompt 를 공유하여 훈련하도록 구현되어 있다. 그러나 이는 서로 다른 특징을 갖는 테스크의 성질이 충분히 반영되기엔 한계가 있는 접근법이다. 따라서 prompt 를 갖고 오는 function의 입력값을 추가하거나 함수자체를 변경하여 각 task 에 적합한 prompt 를 갖고 오도록 하여 인코더 및 태스크 헤드에 입력할 수 있는 접근법을 생각할 수 있다. 만약 학습 시, 같은 종류의prompt를 갖고 온다면 학습이 끝난 후 해당 prompt의 weight 값을 적절한 비율로 혼합해서 prompt pool에 재 저장하는 방식으로 접근할 수도 있다.

Reference:

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

[2] Jia, M., Tang, L., Chen, B., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. (2022). Visual Prompt Tuning. *ArXiv*, abs/2203.12119. <https://doi.org/10.48550/arXiv.2203.12119>.

[3] Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J., Keutzer, K., & Darrell, T. (2022). Multitask Vision-Language Prompt Tuning. *ArXiv*, abs/2211.11720. <https://doi.org/10.48550/arXiv.2211.11720>.

[4] Liu, Y., Lu, Y., Liu, H., An, Y., Xu, Z., Yao, Z., ... & Gui, C. (2023). Hiera rchical Prompt Learning for Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10888-10898).

[5] Wang, Z., Panda, R., Karlinsky, L., Feris, R., Sun, H., & Kim, Y. (2023). Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861*.

[6] Wang, Z., Zhang, Z., Lee, C. Y., Zhang, H., Sun, R., Ren, X., ... & Pfister, T. (2022). Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 139-149).

[7] Bachmann, R., Mizrahi, D., Atanov, A., & Zamir, A. (2022). MultiMAE: Multi-modal Multi-task Masked Autoencoders. *ArXiv*, abs/2204.01678. <https://doi.org/10.48550/arXiv.2204.01678>.

[8] Jacob, G., Agarwal, V., & Stenger, B. (2023). Online Knowledge Distillation for Multi-task Learning. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2358-2367. <https://doi.org/10.1109/WACV56688.2023.00239>.

[9] Rene ́ Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vi- sion transformers for dense prediction. *2021 IEEE/CVF In- ternational Conference on Computer Vision (ICCV)*, pages 12159–12168, 2021. 5, 17

[10] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Fed- erico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239– 248. IEEE, 2016.