Dynamic Inference with Grounding Based Vision and Language Models

Burak Uzkent, Amanmeet Garg, Wentao Zhu, Keval Doshi, Jingru Yi, Xiaolong Wang, Mohamed Omar
Amazon Prime Video
{burauzke,amanmega,zhuwent,kcdos,jyijingr,xiaowanf,omarmk}@amazon.com

1. Experiments

1.1. Implementation Details

1.1.1 MDETR and GLIP

We compare our dynamic inference model, D-ViTMDETR, to both MDETR and GLIP models in large and model small model setups. For the large model set up, for GLIP we use the Swin-B [7] vision transformer with 88M parameters whereas in GLIP [5] Swin-T and Swin-L vision transformers with 29M and 197M parameters are used. For MDETR, we use the ResNet152 [2] backbone with 60M parameters to get large MDETR model together with Roberta-Base text backbone (125M parameters) and multimodal network, DETR, with 18M parameters. For the small model setup for MDETR, we use ResNet101 vision backbone with 22M parameters and CLIP text backbone with 40M parameters whereas we use a DETR architecture with 18M parameters for the multimodal network.

1.1.2 ViTMDETR and D-ViTMDETR

Vision Transformer To utilize ImageNet pre-trained weights, we can use vision transformers pre-trained on either 224×224 pixels or 384×384 pixels images that is available in *timm* library [1]. To achieve higher accuracy, we use a vision transformer, DeiT [9], pre-trained on 384×384 pixels. In our pre-training and finetuning steps, we use 384×384 pixels images in both training and test time. On the other hand, both GLIP and MDETR models use 800×1333 pixels images in test time whereas in training time they use images with different sizes.
Text Transformer For the text transformer, we use the pre-trained Roberta-Base [6] model with 125M parameters for the experiments with large models. For the experiments with small models, we use a customized CLIP model [8] with 40M parameters. We note that this model is not pre-trained.
Multimodal Transformer For processing multimodal representations, we follow MDETR [3] and use the DETR [1] architecture with 6 encoders and 6 decoders that leads to $\sim$ 17M parameters network.
Decision Networks To parameterize the decision networks, we use a single linear layer. For the input to the decision networks, we use the concatenation of the class token embeddings from both the vision and text backbone. The decision network then outputs continuous predictions for the desired number of actions.
Training Hyperparameters For pre-training ViTMDETR, we use the batch size of 256 on 8 NVIDIA V100 GPUs. For the transfer learning tasks for ViTMDETR we use a batch size of 8 with 2 NVIDIA V100 GPUs. For D-ViTMDETR, we use batch size of 256 with 8 V100 GPUs in the transfer learning tasks. We note that our dynamic inference method benefits from large training batch size as it reduces the variance in the reward objective. For the pre-training and finetuning steps, we use the same learning rate and optimization algorithm with MDETR model to pre-train and finetune ViTMDETR model. For D-ViTMDETR model, we use the learning rates of 1e-4 for the decision networks in both pre-training and finetuning steps of the decision networks together with ADAM optimizer [4]. In the joint finetuning step for the decision network and backbones and multimodal network, we use the same learning rates for backbones and multimodal network.
Reward Function Hyperparameter An important hyperparameter in our D-ViTMDETR model is the coefficient, σ, that adjusts the trade-off between computational savings and accuracy of the dynamic inference. With better performing base model (ViTMDETR), we use lower σ value to pay more attention to computational savings. For this reason, for RefCOCO, we set it to 1 whereas for RefCOCOg and RefCOCO+ we set it to 0.8 and 0.6. On the other hand, for GQA, and PhraseCut we set it to 0.4.

1.2. Qualitative Results

In Figure 1, we show some of the predictions of our model on three different group of input pairs. We note that our model allocates smallest amount of resources for the top row, and largest amount of resources for the bottom row, and mid-size amount of resources for the middle row. We can observe that the more complicated the scene

[1] https://github.com/rwightman/pytorch-image-models

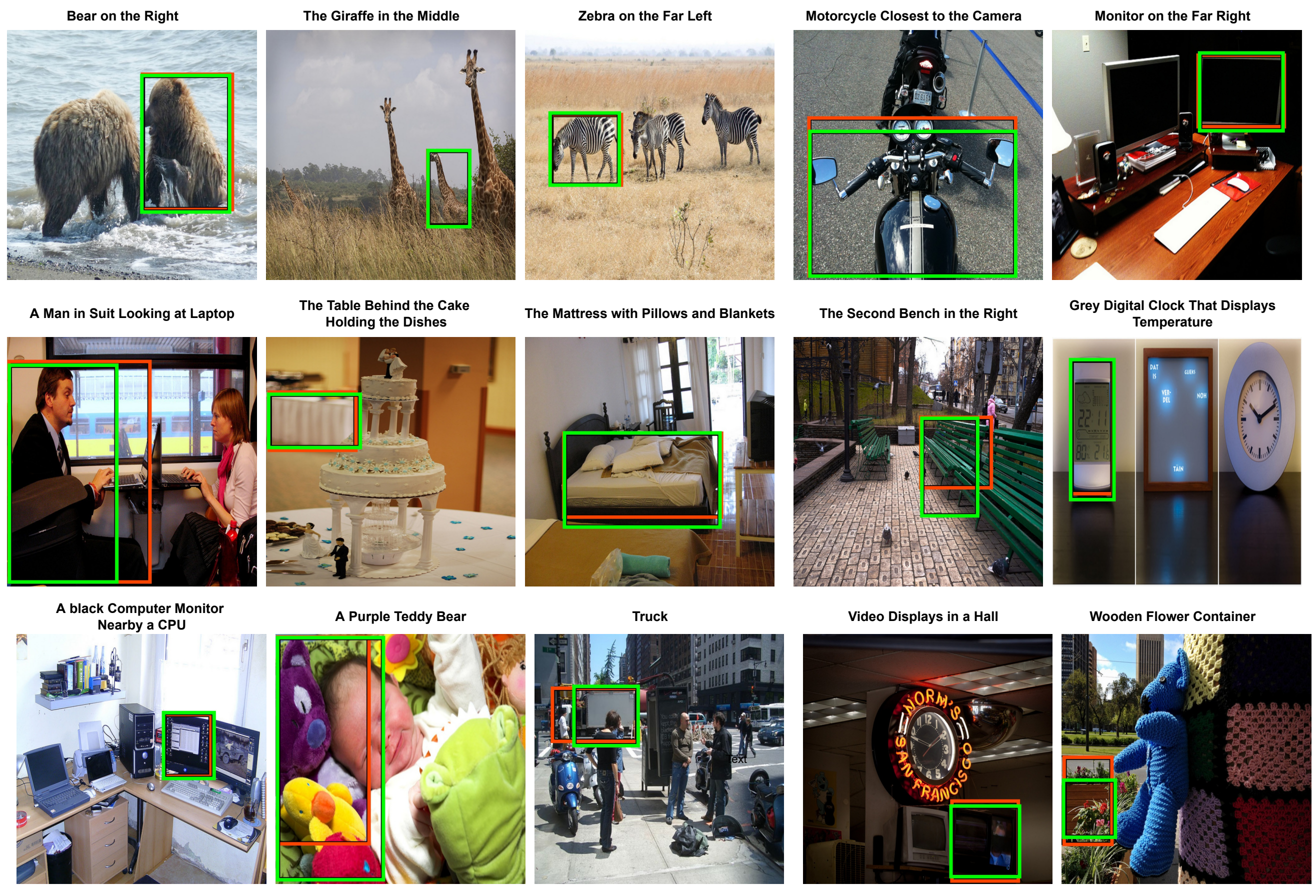

Figure 1. Qualitative results grouped w.r.t the allocated resources by the decision networks. **Top**, **Middle**, and **Bottom** represent smallest, mid-size, and larger amount of allocated resources. Green and red bounding boxes represent the ground truth and predictions.

becomes the more resources are allocated by the decision network.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1