Fine-tuning Vision Transformer-Based Model for Pose-Estimation

Jinxuan Liang, Yihua Zhou, Abdullah Azhar, Anjana Manjunath

Overview

Objective: Fine-Tune the Image-to-text caption generation model, BLIP, for Pose Estimation.

Background: Pose Estimation refers to the task of localizing specific body parts in visual media, and encoding the spatial information into a caption relating body joints with pixel coordinates. The task has a wide range of applications in robotics, virtual reality, and developing accessibility tech.

Data Set: MPII

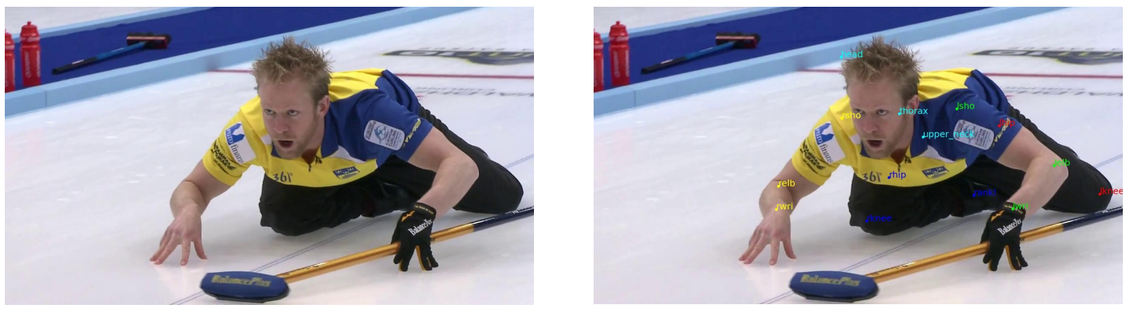

Fig.1: Sample image and overlaid label from the MPII (Andriluka et. al, 2014) . The input image is on the left, and the annotated image is on the right.

- For our study, we only use images in the MPII dataset with one person and all sixteen key points visible.
- This subset results in 1463 images, and with a 80/20 training/validation split, we have 1170 training and 293 validation images.

Methodology

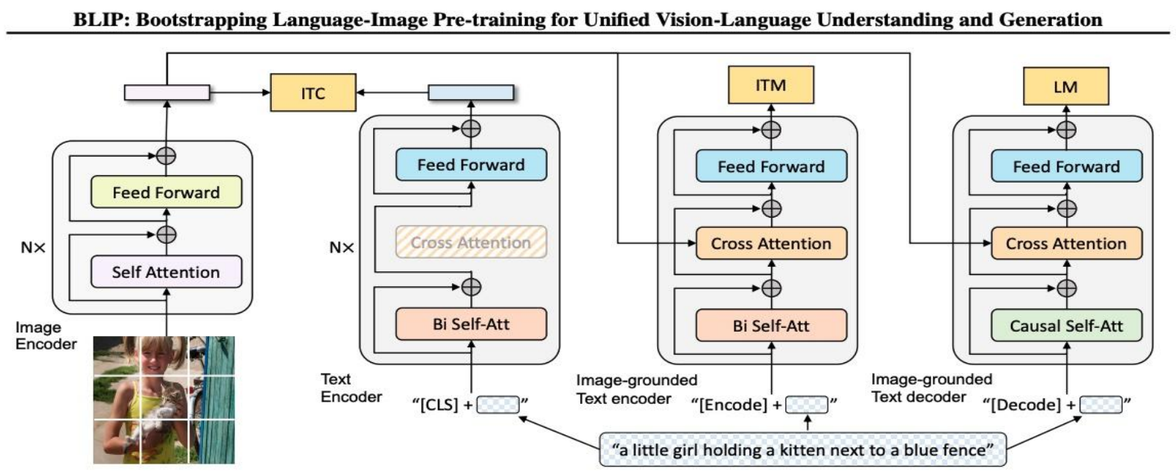

Fig.2: BLIP Architecture from original paper (Li et. al, (2022))

Input:	MPII Dataset Images
Output:	Positional coordinates of joints (e.g., neck (100,200))

- Training Loss: Language Modelling Loss from Text Decoder
- Primary Evaluation Metric: Mean Absolute Error (MAE)
- Validation Threshold Range: [1, 5 , 25] pixels

We fine-tuned BLIP with respect to the hyperparameters listed in the table below. Due to the lack of literature on using BLIP for Pose Estimation (PE) and consequently a lack of prior knowledge on a hyperparameter 'baseline' for fine-tuning BLIP for PE , we used the generic BLIP baselines.

The values in **bold** were determined to be the optimal choices.

Hyper-parameter	Choice of Values
Batch Size	[1, 2, **4**]
Learning Rate	5e−6, **2e−5**, 5e−5
Optimizer	Adam, **AdamW**

Training/Validation

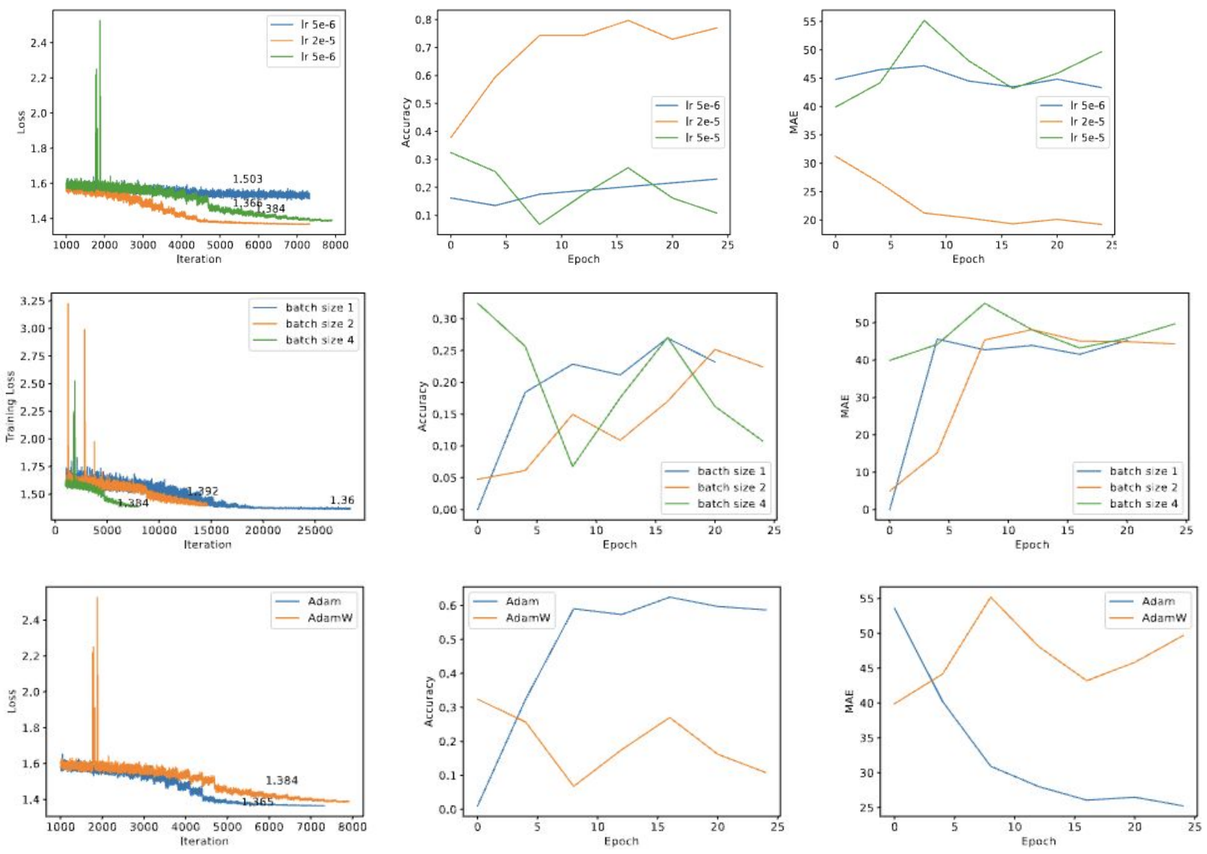

Fig.3: Training Loss, Validation Accuracy and MAE plots for different hyper-parameter combinations.

The validation accuracy on the final model with the optimal hyperparameters was better than anticipated.

Threshold (pixels)	Validation Accuracy
25	**0.925**
5	**0.816**
1	**0.809**

Analysis: Why it works?

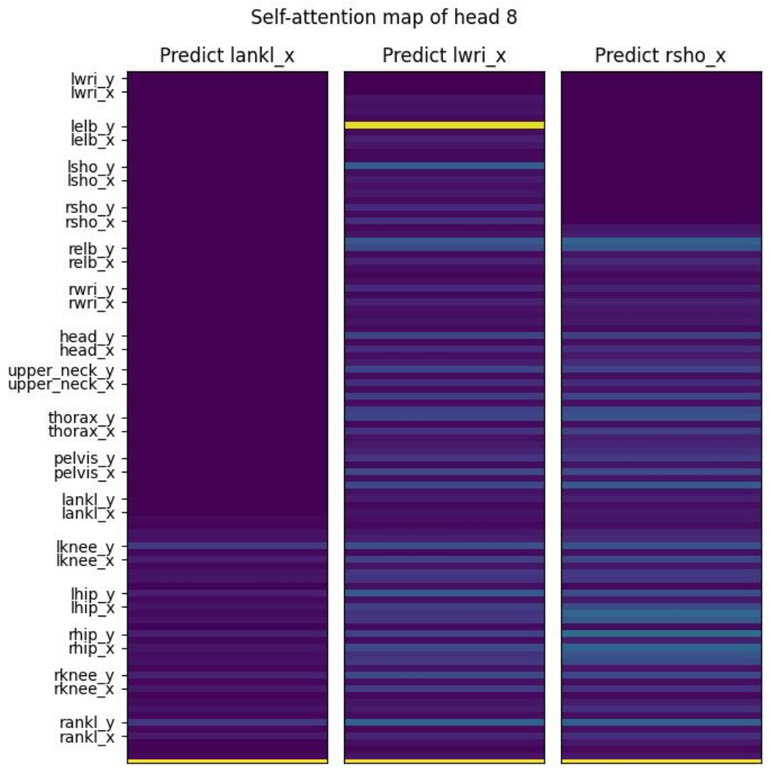

Fig. 4: Self Attention map from attention-head 8 of the text decoder.

- Self attention seems to help the model learn the spatial restrictions between multiple body joints.
- This mechanism may provide the model some flexibility conditional on key points available.
  - Tolerance to occlusion (Fig. 5)

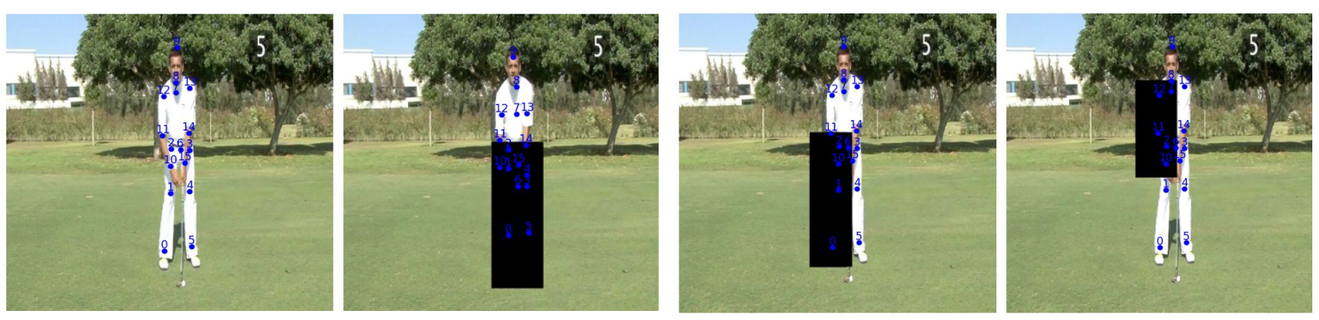

Fig.5: Example of how self-attention mechanism behaves when localizing key points with masking.

Robustness: Images from the Internet

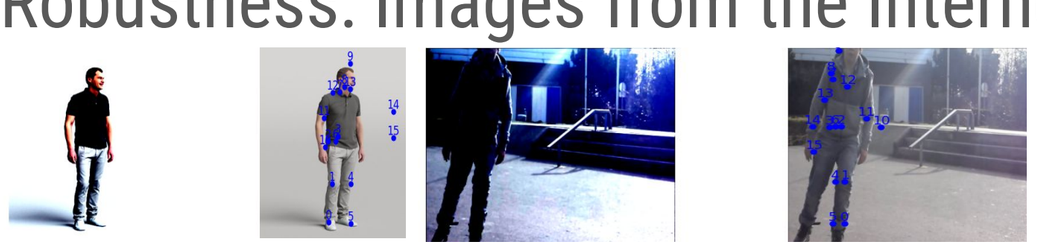

Fig.6: Model outputs on 'out-of-distribution' data

A fine-tuned BLIP has the ability to generalize when generating captions for images from the Internet (out-of-distribution).