Student Engagement Detection in Online Learning Environments using Vision-Language Models and Preference Fine-tuning

Anonymous ECCV 2024 Submission

Paper ID #*****

Abstract. Online learning has been on the rise since the COVID-19 pandemic as it provides the benefits of flexible scheduling, geographic convenience, accessibility, and safer learning environments compared to traditional in-person learning. However, online learning also has disadvantages. One of the major concerns for online learning is whether the online learning environment is able to maintain student engagement and let students focus on the screen during the period of class. Student engagement detection become an important indicator to help teachers to improve student engagement levels in the online learning environment. Vision-language models (VLMs) have given a favorable performance on various computer vision tasks such as recognition, object detection, and tracking. In this paper, we present a method that fine-tunes the vision feature knowledge of the pretrained VLM to provide student engagement descriptions that detect student engagement levels. In addition, we also propose a method to improve the VLM generalizability by performing Direct Preference Optimization(DPO) using self-generate preference pairs. Self-generate pairs allow DPO to provide direct optimization to the broad distribution of generated answers during the fine-tuning process. We compare our student detection method against several vision models and show the superiority of our method. We also show that using DPO with self-generate pairs performs better in terms of generalizability.

Keywords: Student Engagement · Vision-language Model · Direct Preference Optimization

1 Introduction

With the development of network and multimedia technology, online learning environments have been available in daily education. Online learning has been widely used in the education industry due to the various benefits of the online learning environment. For example, online learning offers the opportunity to complete the course anywhere and sometimes even anytime as long as the student resides in a technology-enabled environment. Student engagement refers to the level of the students' focus and effort that they show during the class. Research has demonstrated that student engagement provides a positive influence on academic achievement [3]. Keeping track of the student engagement level is

beneficial for instructors since they can modify their lecture or instruction accordingly to keep the student's attention. In the online learning environment, the monitoring of student engagement level become even more important since it becomes more difficult for instructors to manually track the student engagement level. Therefore, automatic student engagement detection is necessary for the online learning environment.

One popular approach to detecting student engagement is learning the student's facial features and thus obtaining student engagement. Naturally, various vision models have been used to detect student engagement. Plenty of researchers adopted emotion recognition methods through facial expressions for student engagement detection. Then these emotions have been used as cues for different student engagement levels.

2 Background

2.1 Vision-Lauguage Models

Vision-Language Models (VLMs) are multimodal models that put visual and textual information together and can understand the information from both visual and textual channels. There are two major avenues of research on VLMs. One avenue is describing the visual information from the visual knowledge learned by VLMs, to make VLMs comprehend visual knowledge and output textual description. The other is outputting a relevant image from the input textual information. In this paper, we focus on the first avenue.

There are commonly three modules of the VLMs. One vision module, one text module, and a fusion model that associate these two modalities. The vision module and text module are responsible for extracting essential features from the input image and text respectively. The fusion model then closely connects these features together through certain learning strategies. Techniques for building VLMs have evolved over time. The techniques used in recent studies on VLMs are mainly transformer-based techniques. Instead of using hand-designed feature descriptors or pre-trained word vectors/embeddings, transformer-based image and text encoders are employed in VLMs to learn image and text features individually or jointly. One could pretrain the transformer-based models to learn the cross-modal representations first, then these pretrained foundation models could be finetuned for specific downstream tasks.

There are various techniques that have been shown to get good performance and can learn the complicated connections between different modalities. CLIP [], ALIGN [], and DeCLIP [] are vision-language models that bridge the visual and textual modalities by jointly training a text encoder and an image encoder using contractive learning. Contractive Learning computes the distances between image and text data instances, where the distances of matching image-text pairs get minimized distance, and the distances of the image-text pairs that are not matched get maximized.

Similarly to CLIP, BLIP [] uses an image encoder and text encoder to extract visual and textual features. One novelty of BLIP is that it combines three losses

to train the model jointly. Besides the image-text contractive loss (ITC), BLIP uses image-text matching (ITM) loss to learn the difference between positive and negative image-text pairs. A language modeling (LM) loss is used to generate captions for given images. To better associate visual and textual features, the cross attention [] module is used to mix the visual features with the text features sufficiently.

PrefixLM is another learning technique that is mostly used in VLMs. PrefixLM learns to predict the next word in the text sequence given the input part of the text as the prefix. In VLMs, it inputs an image and part of the text and learns to predict the text sequence of words. It applies the same prefix concept to the image by utilizing a Vision Transformer (ViT) [] that divides each image into patches and inputs the sequence of image patches in order into the model. The model obtains the visual embeddings by applying the convolution or linear projection over the image patches. Then, both the image embeddings and the text token embeddings which are converted from the text prefix are inputted into the transformer blocks. SimVLM [] carries out such an architecture using the PrefixLM learning technique.

In addition, learning image embeddings that align with the frozen language model has proved to be more effective in learning a new language task with only a few examples. That is, instead of learning both visual and textual modules, the model now uses a pretrained language model without finetuning and only learns how to align the visual features with the text token embeddings by updating the parameters of the image encoder. There are several VLMs that employ such architecture, such as Frozen [] and ClipCap [].

Furthermore, Flamingo [] freezes both the pretrained vision encoder and the LLM. Flamingo adds new cross-attention layers between the frozen language model layer to tune the language model layer based on the visual input. The perceiver resampler module is added to the frozen vision encoder to produce fewer visual tokens so that the computational complexity of the vision-text cross-attention is reduced.

BLIP-2 [] uses Q-Former, which is a lightweight transformer, to bridge the pretrained frozen image encoders and pretrained frozen LLMs. BLIP-2 uses two stage pretraning for Q-Former. The first stage pretraining learns the vision-language representations. Q-Former sets the learnable embeddings query which extracts visual features (most relevant to the corresponding text) from the input image. Like BLIP, BLIP-2 sets three optimization objectives, image-text matching, image-grounded text generation, and image-text contractive learning. To prevent information leaks, different attention mask is employed for different objectives. The second stage pretraining learns to output visual representations that can be interpreted by the frozen LLM. A fully-connected layer is used to convey the Q-Former output to the same dimension of the chosen LLM input.

2.2 RLAIF

I don't think this should be changed from RLHF to RLAIF

Pretraining The first stage of Reinforcement Learning from Human Feedback(RLHF) is the pretainin stage which trains the language model to predict the next token given the prior text information using the large collection of datasets for natural language processing tasks. The loss used in pretraining stage is normally the cross entropy loss.

Supervised Fine-tuning(SFT) SFT stage finetunes the pre-trained language model on datasets for specific downstream tasks. The datasets are normally high-quality datasets with appropriate instructions (prompts) and reasonable responses. The model obtained after the SFT finetuning is demoted as π_{ref}.

Preference sampling and Reward Learning For a dataset D, for each input x, there is a pair of preference answers (y_ω, y_ι), where y_ω denote the preferred answer human labelers expressed preference for and y_iota denote the dispreferred answer. There is a latent reward model r^* which generates the underlying preferences. A commonly used approach to model preferences is assuming the probability of y_ω is preferred over y_ι as a sigmoid of the reward difference.

$$p^*(y_\omega \succ y_\iota|x) = \sigma(r^*(x, y_\omega) - r^*(x, y_\iota)) \tag{1}$$

A reward model r_ϕ is used to estimate the true reward of human preference via maximum likelihood since the true reward function is not accessible. This is accomplished through minimizing the negative log-likelihood loss.

$$\mathcal{L}_R(r_\phi, D) = -\mathbb{E}_{x,y_\omega,y_\iota\sim D}[\log \sigma(r_\phi(x, y_\omega) - r_\phi(x, y_\iota))] \tag{2}$$

Reinforcement Learning Optimization The RL phase uses the learned reward function to further optimize the language model. To prevent model collapse and maintain the proximity to the distribution of reward model π_{ref}, a KL divergence penalty is added.

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}[r_\phi(x, y)] - \beta\mathbb{D}_{KL}[\pi_\theta(y|x) \parallel \pi_{ref}(y|x)] \tag{3}$$

Where π_θ is the model that RLHF is optimizing, β is the hyperparameter to restrict on how far the model can deviate from the base reference model. Due to the non-differentiability, the objective is normally optimized with an RL algorithm such as PPO [].

2.3 Multi-Modal Preferenec Alignment

[4] [5] [6] [1]

3 Methodology

We propose a new method to detect the engagement level of students. We utilize the description generation abilities of vision-language models. In this paper, we build our model based on MiniGPT-4 [**?**] which aligns visual information from a frozen vision encoder with a frozen advanced large language model (LLM) [] using one projection layer. We finetune MiniGPT-4 on our dataset using a set of prompts that are more relevant to the student engagement problem. In order to achieve more general performance, we optimize the MiniGPT-4 model with Direct Preference Optimization (DPO) [] with self-generate preference pairs.

3.1 Finetuning MiniGPT-4 with Student Engagement Dataset

The vision encoder employed in MiniGPT-4 consists of a ViT [] backbone and their pre-trained Q-Former []. For the language decoder, MiniGPT-4 utilized the Vicuna [] and llama2 model which is constructed upon LLaMA.

MiniGPT-4 adopted a two-stage training process. The **first stage** is the pretaining stage which aims to obtain the vision language knowledge. The dataset used in the pretaining stage is a combined large collection of the aligned image-text dataset of Conceptual Caption, SBU, and LAION. However, after this stage, the outputs that are generated may contain incoherent linguistic outputs, such as repetitive words or sentences, fragmented sentences, or irrelevant content.

The **second stage** alignment process utilizes the image-text pairs, in which the image descriptions are automatically generated by MiniGPT-4 and then mended by ChatGPT, to finetune the pretrained model of the first stage. A prompt that follows the Vicuna language model is added to encourage MiniGPT-4 to generate more detailed descriptions to form image-text pairs. The prompt is shown as follows:

###Human: <Img><ImageFeature></Img>Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:

###Human: Continue ###Assistant: is also used as an additional prompt if the generated text is less than 80 tokens. This prompt would let MiniGPT-4 extend the text generation process. Then a post-processing is performed to fix the noisy or incoherent descriptions that are possible to happen in the MiniGPT-4 model from the first pretraining stage. The post-processing includes using ChatGPT to mend the descriptions and manually verify if the descriptions are correct for the corresponding images.

After the post-processing of image-text pairs, MiniGPT-4 is finetuned with the prompts follow the templates shown below:

###Human: <Img><ImageFeature><Img><Instruction>###Assistant: The $< Instruction >$ is randomly sampled from a predefined instruction set. This instruction set contains different forms of instructions as shown below:

- *Describe this image in detail.*
- *Take a look at this image and describe what you notice.*
- *Please provide a detailed description of the picture.*

– *Could you describe the contents of this image for me?*

After the second stage, The output descriptions generated by MiniGPT-4 of the corresponding images becomes more natural, reliable, and human-friendly compared to the first stage.

Stundent Engagement Detection Finetuning In our model, we introduce a different finetuning that aims to generate specific image descriptions to describe the different student behaviors which indicate different engagement levels of student. In this paper, we use the Student Engagement Dataset (SED) [2] to finetune the MiniGPT-4 model. There are three different categories in SED, 'looking at their paper', 'looking at their screen', or 'wandering'. According to these different labels, we set the corresponding prompts as follows:

– *The person is looking down at the paper*
– *The person is looking straight at the screen*
– *The person is looking away*

Accordingly, we modify the instruction set of prompts from the general instruction to the questions that are used specifically to determine the students' behavior. The questions used in our task are shown below:

– *Is the person looking straight at the screen?*
– *Is the person looking down at the paper?*
– *Is the person looking away?*
– *Is the person looking straight at the screen? Is the person looking down at the paper? Is the person looking away?*

The loss we used for the finetuning is the same as the LLaMA model. We calculate the cross entropy loss between the tokens generated by our model and the labels that we targeted. We shift the generated tokens and the targeted labels so that the model predicts the next token based on the previous one.

3.2 Direct Preference Optimization using Self-generate Preference Pairs

Direct Preference Optimization (DPO) provides a simplification of traditional feedback-aligned LLMs. DPO avoids training a Reinforcement Learning (RL) loop as RLHF. Instead, DPO proposes a closed-form loss that leverages a particular choice of reward model parameterization. Equation 4 shows the detail of DPO loss, where y_ω represents the preferred answer and y_ι represents the dispreferred answer. DPO loss updatings mainly maximize the margin between preferred and dispreferred answer pairs as shown in equation 5.

$$L_{DPO}(\pi_\theta, \pi_{ref}) = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_\omega|x)}{\pi_{ref}(y_\omega|x)} - \beta \log \frac{\pi_\theta(y_\iota|x)}{\pi_{ref}(y_\iota|x)}\right)\right] \quad (4)$$

$$r_\omega = \log(\pi_\theta(y_\omega|x)) - \log(\pi_{ref}(y_\omega|x))$$

$$r_\iota = \log(\pi_\theta(y_\iota|x)) - \log(\pi_{ref}(y_\iota|x))$$

$$L_{DPO}(\pi_\theta, \pi_{ref}) = -\mathbb{E}[\log \sigma(\beta(r_\omega - r_\iota))] \quad (5)$$

The datasets that are used in academic research areas of feedback-aligned language models are preference data pairs such as HH, SHP, and OASST. However, for the student engagement detection problem, we normally work with multi-class labeled datasets. Various labels in the datasets indicate the different engagement behaviors of the student. Therefore, we need to transform the labeled datasets to preference dataset $D = \{x^i, y^i_\omega, y^i_\iota\}_{i=1}^N$, where N denote the number of data samples where for each input image x, we assign an answer pair of preferred answer and dispreferred answer (y_ω, y_ι). A simple approach is to utilize the given ground truth to generate a preference dataset by using the given label of each image the preferred answer and using the other labels in the dataset as the dispreferred answer. However, this approach will introduce bias to the vision-language model since the distribution of possible dispreferred answers has a much wider range.

In order to yield better answers, it is reasonable to extend the reference data pairs D so that the dispreferred answer comes from a more general distribution than ground truth. Since we do not have the distribution of generated answers of the vision-language model, we use the answers generated from the vision-language model itself to extend the dispreferred answers. In practice, we form the dispreferred answers during training in the following manner. For $a\%$ of the time we use the generated wrong answers as the disprefered answers. For $(100 - a)\%$ of the time, we randomly choose from other labels besides the correct label as the dispreferred answers.

4 Evaluation

4.1 Dataset

The dataset used in this paper is the Student Engagement Dataset(SED) [2]. This dataset contains a raw dataset and a balanced dataset. The raw dataset contains 18,721 frames which are sampled at one frame-per-second(FPS) from 400 videos collected from 19 students. The balanced dataset is a smaller version that removes similar samples for each class and balances the raw dataset. The balanced dataset contains 1973 frames. Both the raw dataset and balanced dataset contain three categories of student data, paper, screen, and wander.

DAiSEE [] dataset is a large labeled student engagement level dataset which is collected by a web camera during the period of student watching educational and recreational videos. The dataset contains 9068 video snippets collected from 32 female and 80 male subjects of age 18 to 30. There are 6 differnet location

settings (dorm rooms, crowded lab spaces, library etc) and 3 different illumination settings (light, dark and neutral) of the video environment in DAiSEE dataset. The dataset is original labeled with 4 different student engagement levels: boredom, confusion, engagement, frustration. In our evaluation, we sampled the videos to various frames and randomly selected 1046 frames out of it. Then we relabeled these frames to the previous three categories (paper, screen, wander). The idea is to evaluate whether the model can apply the knowledge learned from the student engagement dataset to a unknown dataset.

Preprocessing We used 80% of the data samples from the balanced SED dataset for finetuning and 20% of the samples for our evaluation with balanced data. We also evaluated our model with raw SED dataset. We exclude the training balanced sample from the raw SED dataset and used the rest of the samples for the evaluation. For both training and testing data samples, we combined the labels for the various categories and created an annotation JSON file that contains the image ID and the corresponding descriptions. We also compare the results of our finetuning prompts with the original prompts used by MiniGPT-4. The results of evaluation on original MiniGPT-4 checkpoint without finetuning on SED dataset is also shown as the baseline of MiniGPT-4 model.

We evaluated two different versions of MiniGPT-4 in this paper. One version is that we perform our finetuning using the MiniGPT-4 (Vicuna) checkpoint. Another version is that we perform our finetuning after using the MiniGPT-4 (Llama2) checkpoint. The MiniGPT-4 (Vicuna) achieve an accurary of 96.2% after our finetuning while MiniGPT-4(Llama2) gives a evaluation result of 88.6%. Therefore, we use MiniGPT-4 (Vicuna) as our base model.

For the evaluation on DAiSEE dataset, we manually labeled 1046 frames and assigned image ID for each frame and constructed a annotation JSON file with the same format of the SED dataset.

4.2 Results

During the evaluation, we used the same prompt to ask the model to answer "Is the person looking straight at the screen? Is the person looking down at the paper? Is the person looking away?" for all the samples. If the generated sentence contains the desired image description, we consider it as the correct answer. We evaluate the accuracy of our model using the f1 score and compare it with the original deep learning vision model results. As stated previous, we finetuned our model on the balanced SED dataset and evaluated both the balanced dataset and the raw SED dataset. We show the evaluation results on SED balanced dataset in table table 1. As we can see from this table, our method achieved 96.2% accuracy compared to the original deep learning vision models. We show the evaluation results on SED raw dataset in table 2. Table 3 shows the evaluation results on of DAiSEE dataset.

Method	Accuracy (%)
MiniGPT-4 & DPO	96.2
MiniGPT-4 (SED specific prompts)	96.2
MiniGPT-4 (Original prompt)	89
MobileNet (Pretrained ImageNet)	94
Xception (Pretrained ImageNet)	88
VGG16 (Pretraiend ImageNet)	85
Head Pose Estimator (Logistic Reg.)	60
Head Pose Estimator (Conditional)	55
MiniGPT-4 (Original model)	54.3

Table 1: Accuracy of student engagement detection on SED balanced dataset.

Method	Accuracy (%)
MiniGPT-4 & DPO	94.1
MiniGPT-4 (SED specific prompts)	93.1
MiniGPT-4 (Original prompt)	88.6
MiniGPT-4 (Original model)	58.2

Table 2: Accuracy of student engagement detection on SED raw dataset

Method	Accuracy (%)
MiniGPT-4 & DPO	86.9
MiniGPT-4 (SED specific prompts)	86.5
MiniGPT-4 (Original prompt)	85.6
MiniGPT-4 (Original model)	50.9

Table 3: Accuracy of student engagement detection on DAiSEE dataset

Method	Accuracy (%)
MiniGPT-4 & DPO	84.7
MiniGPT-4 (SED specific prompts)	82.4
MiniGPT-4 (Original prompt)	70.6
MiniGPT-4 (Original model)	27.1
ChatGPT4	74.2

Table 4: Accuracy of student engagement detection on SED handpicked dataset

5 Related Works

6 Conclusion

Page 10 of the manuscript.

Page 11 of the manuscript.

Page 12 of the manuscript.

Page 13 of the manuscript.

Page 14 of the manuscript. This is the last page.

Now we have reached the maximum length of an ECCV 2024 submission (excluding references). References should start immediately after the main text, but can continue past p. 14 if needed.

References

1. Bhatia, G., Nagoudi, E.M.B., Cavusoglu, H., Abdul-Mageed, M.: Fintral: A family of gpt-4 level multimodal financial large language models (2024) 4
2. Delgado, K., Origgi, J.M., Hasanpoor, T., Yu, H., Allessio, D., Arroyo, I., Lee, W., Betke, M., Woolf, B., Bargal, S.A.: Student engagement dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3628–3636 (October 2021) 6, 7
3. Lei, H., Cui, Y., Zhou, W.: Relationships between student engagement and academic achievement: A meta-analysis. Social Behavior and Personality: an international journal **46**, 517–528 (03 2018). https://doi.org/10.2224/sbp.7054 1
4. Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L.: Silkie: Preference distillation for large visual language models (2023) 4
5. Li, S., Lin, R., Pei, S.: Multi-modal preference alignment remedies regression of visual instruction tuning on language model (2024) 4
6. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization (2024) 4