T.C.

MARMARA UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

CSE4078 Introduction to NLP

DOWNSTREAM TASK REPORT

May 30th, 2024

Group 8 Members

Alper Özdemir

Eren Başpınar

Emirhan Erdoğan

Faruk Akdemir

Şule Koca

Downstream Task Evaluation

Task Description

The downstream task evaluated in this report is the HellaSwag Turkish (hellaswag_tr) task. This task is a benchmark designed to test natural language understanding and commonsense reasoning capabilities of models. The task involves selecting the most plausible continuation of a given scenario from multiple choices.

The evaluation was conducted under two conditions: zero-shot and one-shot settings. In the zero-shot setting, the model makes predictions without any prior examples, while in the one-shot setting, the model is given one example to guide its predictions. The primary evaluation metrics used are accuracy (acc) and normalized accuracy (acc_norm), along with their respective standard errors (Stderr).

Evaluations

The following models were evaluated:

1. gemma-2b-CSE4078S24_Grp8-r8-4bit-tr (r = 8, lora_alpha = 8)
2. gemma-2b-CSE4078S24_Grp8-r16-4bit-tr (r = 16, lora_alpha = 16)
3. gemma-2b-CSE4078S24_Grp8-r32-4bit-tr (r = 32, lora_alpha = 32)
4. gemma-2b-CSE4078S24_Grp8-r64-4bit-tr (r = 64, lora_alpha = 64)
5. gemma-2b-CSE4078S24_Grp8-r128-4bit-tr (r = 128, lora_alpha = 128)
6. gemma-2b-CSE4078S24_Grp8-r256-4bit-tr (r = 256, lora_alpha = 256)

Evaluation Results

Zero-shot Setting

Model Name	acc	Stderr	acc_norm	Stderr
gemma-2b-CSE4078S24_Grp8-r8-4bit-tr	0.3323	±0.0149	0.3771	±0.0153
gemma-2b-CSE4078S24_Grp8-r16-4bit-tr	0.3313	±0.0149	0.3781	±0.0153
gemma-2b-CSE4078S24_Grp8-r32-4bit-tr	0.3303	±0.0148	0.3751	±0.0153
gemma-2b-CSE4078S24_Grp8-r64-4bit-tr	0.3284	±0.0148	0.3672	±0.0152
gemma-2b-CSE4078S24_Grp8-r128-4bit-tr	0.3144	±0.0147	0.3632	±0.0152
gemma-2b-CSE4078S24_Grp8-r256-4bit-tr	0.3134	±0.0146	0.3493	±0.0150

One-shot Setting

Model Name	acc	Stderr	acc_norm	Stderr
gemma-2b-CSE4078S24_Grp8-r8-4bit-tr	0.3323	±0.0149	0.3642	±0.0152
gemma-2b-CSE4078S24_Grp8-r16-4bit-tr	0.3323	±0.0149	0.3672	±0.0152
gemma-2b-CSE4078S24_Grp8-r32-4bit-tr	0.3294	±0.0148	0.3652	±0.0152
gemma-2b-CSE4078S24_Grp8-r64-4bit-tr	0.3284	±0.0148	0.3662	±0.0152
gemma-2b-CSE4078S24_Grp8-r128-4bit-tr	0.3234	±0.0148	0.3642	±0.0152
gemma-2b-CSE4078S24_Grp8-r256-4bit-tr	0.3134	±0.0146	0.3463	±0.0150

Analysis and Discussion

The evaluation results across the models indicate a modest level of performance in both zero-shot and one-shot settings.

Zero-shot Analysis

- The models show a slight decline in accuracy as the rank (r) increases. This might suggest that increasing the rank beyond a certain point does not yield proportional improvements and might even degrade performance.
- The normalized accuracy values are consistently higher than the raw accuracy, indicating that the models are relatively better at ranking the correct answers.

One-shot Analysis

- The accuracy values in the one-shot setting are quite similar to those in the zero-shot setting, with very little to no improvement observed.
- The normalized accuracy shows a slight increase in the one-shot setting, suggesting a marginal benefit from the additional example.
- Overall, the r8 and r16 models performed slightly better compared to higher rank models. This indicates that a moderate rank and lora_alpha might be optimal for this specific task and dataset.

Conclusion

The evaluation of the GEMMA-2B derived models on the HellaSwag Turkish task demonstrates that the models achieve a baseline performance, with the accuracy values ranging around 0.33. The normalized accuracy values are slightly higher, reflecting the models' ability to rank choices effectively.

The results suggest that increasing the rank (r) and lora_alpha values beyond a certain point does not necessarily improve performance and might even be detrimental. The r8 and r16 models appear to be the most effective in this evaluation, balancing performance and computational efficiency.

Further research could involve fine-tuning the models on task-specific data and experimenting with different configurations to enhance performance. Additionally, exploring other downstream tasks could provide a broader understanding of the models' capabilities and limitations.