Enhancing Numerical Prediction of MLLMs with Soft Labeling

Pei Wang, Zhaowei Cai, Hao Yang, Davide Modolo, Ashwin Swaminathan

Amazon AGI

amazon

Motivation

- MLLMs have achieved remarkable progresses across various tasks.
- The learning is to predict the next token, associated with the *de facto* cross-entropy loss across the token vocabulary, no matter the target is a word, digit, punctuation, or anything else.
- Is the standard cross-entropy loss optimal for any LMMs/MLLMs tasks?

amazon

Numerical Prediction

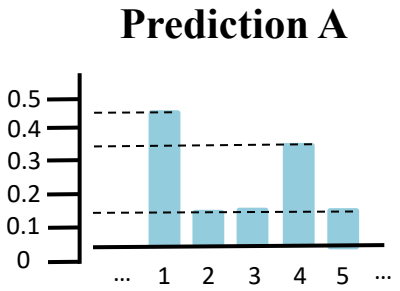
"How many trains in this image?"

amazon

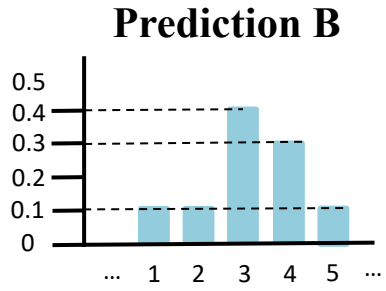
Numerical Prediction

"How many trains in this image?"

MLLM



1 train



3 trains

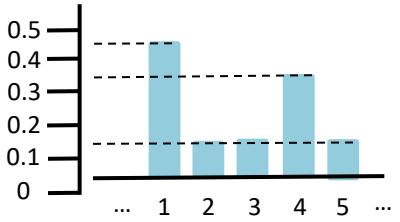
amazon

Numerical Prediction

"How many trains in this image?"

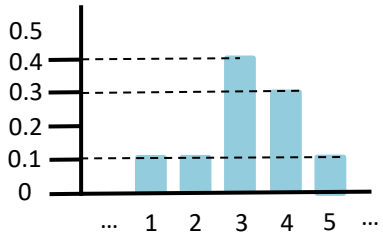
MLLM

Prediction A



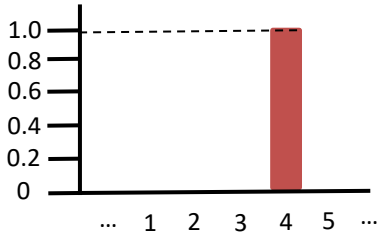
1 train

Prediction B



3 trains

Hard labeling



Loss A=1.20

Loss B=1.20

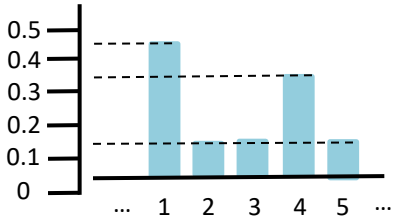
amazon

Numerical Prediction

"How many trains in this image?"

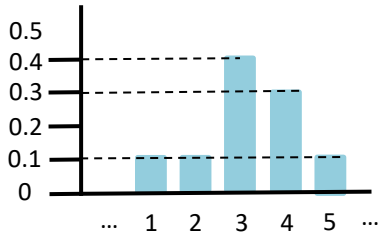
MLLM

Prediction A



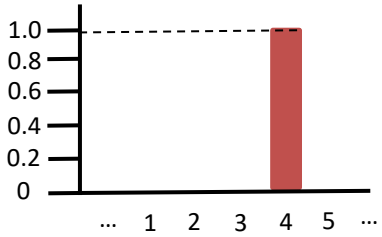
1 train

Prediction B



3 trains

Hard labeling



Loss A=1.20

Loss B=1.20

Loss A = loss B, the real distance between the prediction and target is ignored.

amazon

Cross-entropy Loss

- Unlike discrete tokens (words), in **numerical prediction** the de facto cross-entropy loss with one-hot encoded target fails to account for the distance between predicted and target tokens *properly*.

$$L_{CE}(\mathbf{p}, \mathbf{q}(t)) = \sum_{i=1}^{V} q_i(t)(-\log p_i),$$

amazon

Cross-entropy Loss

- When a prediction is incorrect, we should measure how wrong it is by assessing the distance from the target, rather than treating all wrong predictions equally as long as the confidence on the target token is the same, as is done in hard labeling.

$$L_{CE}(\mathbf{p}, \mathbf{q}(t)) = \sum_{i=1}^{V} q_i(t)(-\log p_i),$$

amazon

Soft Labeling Cross-entropy Loss

- We defines soft label for the 10 digit tokens from “0” to “9” as

$$\mathbf{q}^{SL}(t) = (1-\eta)\boldsymbol{\delta}(t) + \eta\boldsymbol{\psi}(t)$$

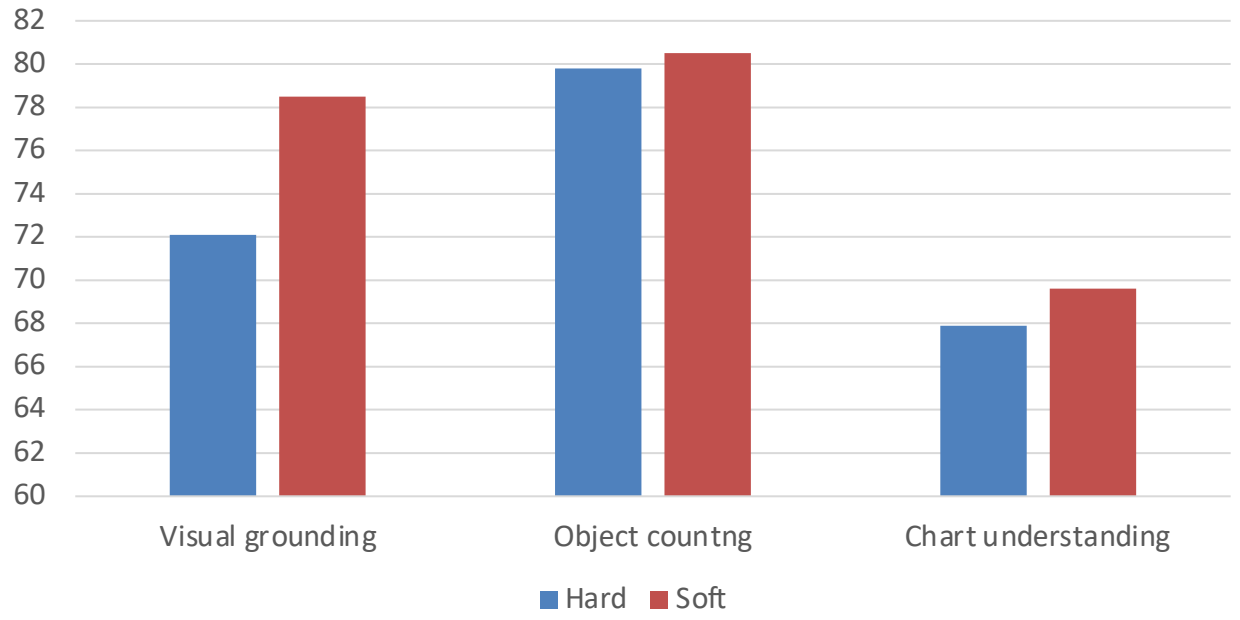
- The final loss is

$$L = \frac{1}{N_r + N_n}\left(\sum_{(x_r,y_r)\in(\mathbf{x}_r,\mathbf{y}_r)} L_{CE}(f(x_r), \mathbf{q}(y_r)) + \lambda \sum_{(x_n,y_n)\in(\mathbf{x}_n,\mathbf{y}_n)} L_{CE}(f(x_n), \mathbf{q}^{SL}(y_n))\right),$$

amazon

Advancement of Soft Labeling

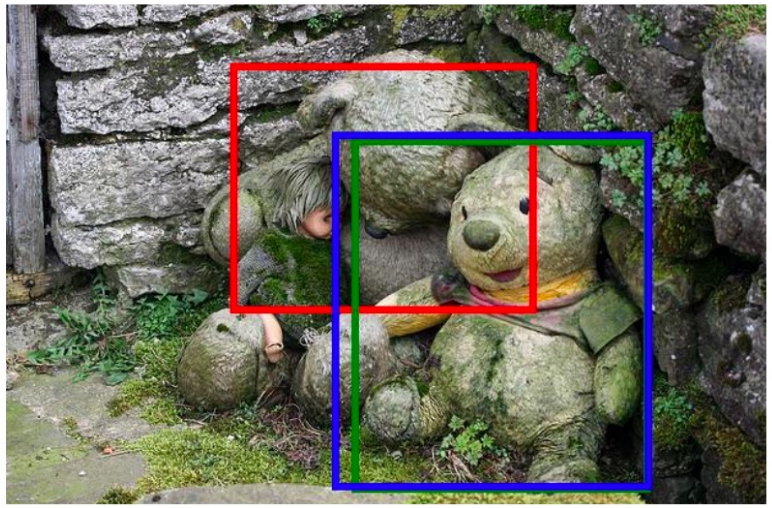
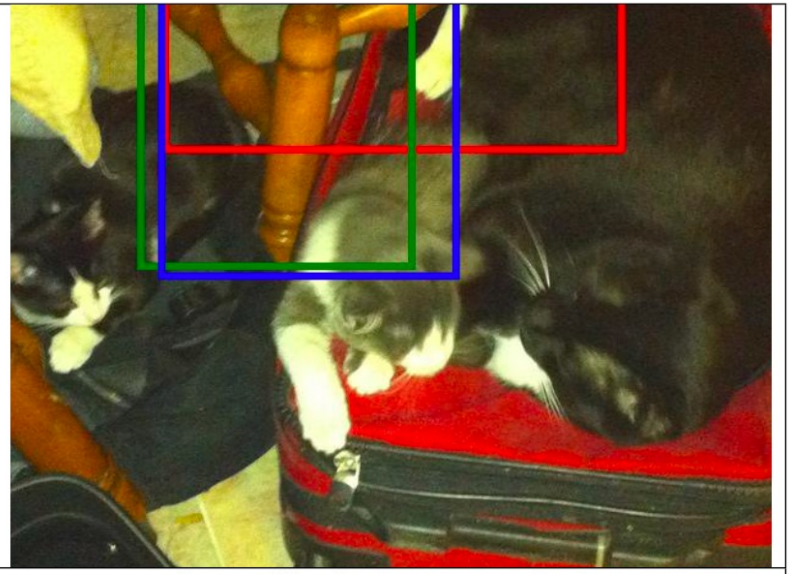
- Effectiveness on visual grounding, object counting, chart understanding



amazon

Advancement of Soft Labeling

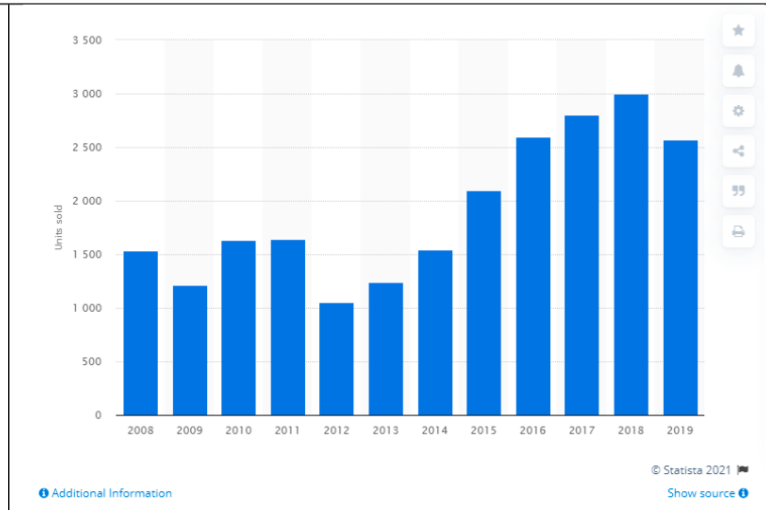
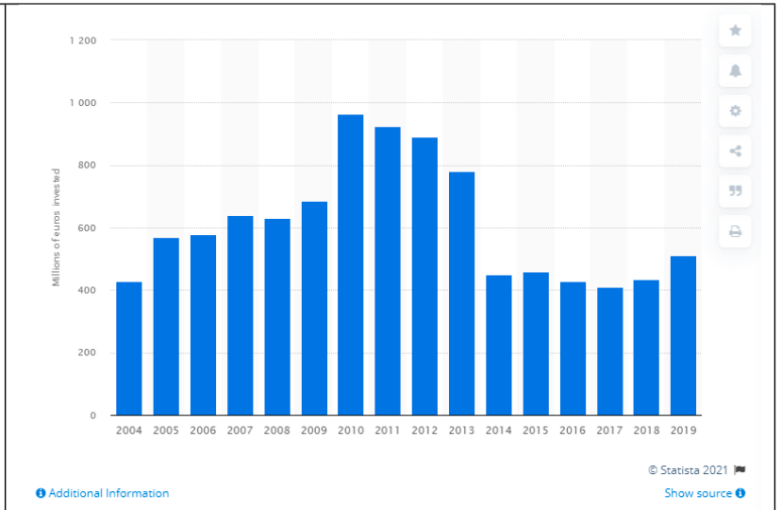
- Effectiveness on visual grounding, object counting, chart understanding

Image		
Prompt	Rocks that look like Winnie-the-poo, facing the camera	wooden stool leg between cats and suitcases
Hard labeling	red box	red box
Soft labeling	green box	green box
Ground truth	blue box	blue box

amazon

Advancement of Soft Labeling

- Effectiveness on visual grounding, object counting, chart understanding

Image		
Prompt	How many cars did Mini sell in Portugal as of 2019?	What was the total amount of investments in sea port infrastructure in 2010?
Hard labeling	2301	800
Soft labeling	2701	930.5
Ground truth	2601	965

amazon