DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding

Wenwen Yu, Zhibo Yang, Yuliang Liu, Xiang Bai

Marco VItali

Università degli Studi di Firenze
Corso di Laurea in Data Science, Calcolo Scientifico & Intelligenza Artificiale
B033612 (B241) - NATURAL LANGUAGE PROCESSING 2025-2026

9 dicembre 2025

Indice

1.1 Contesto

Gli MLLM hanno dimostrato grandi performance nel document understanding, ma hanno ancora certe limitazioni:

- **Black-box:** si ha poca fiducia in ambiti ad alto rischio, come quello medico, finanziario o legale.
- **CoT statiche e SFT:** soffrono di catastrophic forgetting, scarsa adattabilità e limitata generalizzazione.

1.2 Presentazione del Framework e Obiettivi

Per cercare di colmare queste lacune viene presentato **DocThinker**, ovvero un nuovo framework di Rule-Based Reinforcement Learning

- Apprendimento dinamico della policy.
- Vincolo KL.
- Risultati intermedi più compensibili per l'uomo.

Obiettivi:

- Mitigare il catastrophic forgetting.
- Migliorare Adattabilità, Trasparenza e Generalizzazione.

2.1 Tecniche di Ragionamento CoT

- **ReFocus:** Framework basato sulla modifica visiva che migliora la comprensione di tabelle e grafici e introduce la *modifica attiva delle immagini*.
- **Visual CoT:** Framework con passi visivi che introduce una pipeline di elaborazione multi-turn e adotta un metodo di focalizzazione visiva guidata.
- **MVoT:** Genera *tracce di ragionamento visive e testuali intercalate* migliorando la trasparenza ma non la capacità di ragionamento.
- **The Mind's Eye of LLMs:** Sollecita il ragionamento spaziale generando rappresentazioni visive dei processi di pensiero, si applica molto alla navigazione ma ha una generalizzazione molto limitata per il document understanding.

2.2 Reinforcement Learning

- **DeepSeek-R1:** Dimostra che l'apprendimento basato su RL può incentivare comportamenti di ragionamento emergente ottenendo buone prestazioni anche su compiti elaborati.
- **MedVLM-R1:** Applica il RL a modelli visivo-linguisici in medicina e mostra che *migliorano la trasparenza e la generalizzazione*.
- **Visual-RFT:** Introduce il Visual Reinforcement Fine-Tuning migliorando i ragionamenti visivo-linguistici, inoltre mostra miglioramenti significativi nell'efficienza dei dati, nel riconoscimento di oggetti few-shot e nel grounding di ragionamento.

2.3 DocThinker

Nonostante gli evidenti risultati ottenuti applicando il RL alle strategie di ragionamento degli MLLM, questa è un'area ancora abbastanza inesplorata, in particolare per quanto riguarda la progettazione di funzioni di ricompensa efficaci.
Qui entra in gioco DocThinker:

- Funzioni di ricompensa multi-obiettivo.
- Domande riformulate.
- Regions of Interest.
- Apprendimento della policy tramite l'algoritmo GRPO.
- Vincolo KL.
- Consente l'autorevisione e la correzione.

2.4 Lavori Correlati

Esistono altri lavori correlati in cui i modelli MLLM, quali LLaVAR, DocLLM e DocPedia, cercano di portare migliorie nel document understanding, ma si evidenzia una forte necessità di un framework di ragionamento adattivo.
Per questo motivo in altri modelli viene applicato il RL come alternativa e con questi lavori si dimostra il suo successo nelle task visivo-linguistiche, da qui e dalla sopracitata mancanza di esplorazione di questo ambito nasce DocThinker.

3.1 Algoritmo GRPO

Framework di RL che non richiede un modello critic, eliminando costi computazionali e instabilità, confrontando un gruppo di risposte candidate.

- Da una domanda q il modello precedente π_{OLD} genera un gruppo di G risposte candidate $\{O_1,...,O_G\}$
- Si valutano le risposte con funzioni di ricompensa R(q,O) ottenendo le ricompense $\{r_1,...,r_G\}$
- Le ricompense vengono normalizzate creando dei vantaggi $\{A_1,...,A_G\}$ e si pone per ogni token t della sequenza di risposta $A_{i,t} = A_i$

3.1 Algoritmo GRPO

- Si calcola la Loss come

$$L_{GRPO}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|O_i|}\sum_{t=1}^{|O_i|}[\frac{\pi_\theta(O_{i,t}|q,O_{i,<t})}{\varphi[\pi_\theta(O_{i,t}|q,O_{i,<t})]}A_{i,t}+$$

$$-\beta D_{KL}(\pi_\theta||\pi_{ref})]$$

- Backpropagation e aggiornamento dei pesi

3.2 Modello Base e Prompt Template

- Il modello di base utilizzato è Qwen2.5-VL nelle varianti 3B e 7B, un modello linguistico multimodale all'avanguardia.
- Il Prompt Template è strutturato per istruire il modello a fornire una traccia di ragionamento ed una risposta codificati in tag simili a XML del tipo <think>, </think> e <answer>, </answer>.
- La risposta finale è formattata in JSON e contiene 3 campi fondamentali:
  - **rephrase_question**
  - **bbox_2d**
  - **final_answer**

3.3 Funzioni di Ricompensa Multi-Obiettivo

Gli approcci tradizionali si basano su feedback umani, DocThinker utilizza invece il Reinforcement Learning with Verifiable Rewards basandosi su 4 criteri fondamentali:

- **Format Reward** R_{format}
- **Accuracy Reward** $R_{accuracy}$
- **RoI IoU Reward** R_{RoI}
- **Rephrase Question Reward** $R_{rephrase}$

3.3 Funzioni di Ricompensa Multi-Obiettivo

R_{format}, $R_{accuracy}$ e R_{RoI} possono valere 1 nel caso in cui le condizioni imposte sono rispettate o 0 altrimenti, mentre $R_{rephrase} = s + r$ se $R_{accuracy} = 1$ e 0 altrimenti, dove s rappresenta la cosine similarity tra la domanda originale e quella riformulata e r il numero di parole nuove rispetto a quelle totali nella domanda riformulata.
Infine la funzione di ricompensa finale R_{tot} si calcola come

$$R_{tot} = \lambda_1 R_{format} + \lambda_2 R_{accuracy} + \lambda_3 R_{RoI} + \lambda_4 R_{rephrase}$$

ponendo $\lambda_i = 1$ per ogni $i \in \{1; 2; 3; 4\}$

4.1 Dataset e dettagli

Il dataset per l'addestramento è utilizzato è VisualCoT, il quale contiene 438000 coppie domanda-risposta con regioni chiave evidenziate e ricopre diversi domini.
Vengono prese 2 configurazioni di questo dataset:

- **4data4k:** 4x1000 campioni da:
  - DocVQA
  - InfographicsVQA
  - TextCaps
  - TextVQA
- **8data8k:** campioni di 4data4k + 4x1000 campioni da:
  - Flickr30k
  - GQA
  - Open Images
  - VSR

4.1 Dataset e dettagli

Il modello di base viene addestrato con:

- 2 Epoche
- 8 GPU NVIDIA A100 da 80GB
- Batch Size=2
- Numero di Risposte Candidate G=6
- Ottimizzatore AdamW
- Learning Rate 10^{-6}
- Coefficiente KL $\beta = 0.04$

La valutazione avviene secondo il protocollo di valutazione di VisualCoT che prevede anche l'utilizzo di dataset zero-shot, quali SROIE, DUDE e Visual7W e avviene tramite le metriche standard fornite dal VisualCoT Benchmark.

4.2 Risultati Principali

MLLM	Res.	Data	Str.	Document-oriented Understanding						General Multimodal Understanding				
				Doc/Text					Chart	General VQA		Relation Reasoning		
				DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfoQA	F30k	V7W	GQA	OI	VSR
LLaVA-1.5-7B [22]	336^2	-	SFT	0.244	0.597	0.588	0.290	0.136	0.400	0.581	0.575	0.534	0.412	0.572
LLaVA-1.5-13B [22]	336^2	-	SFT	0.268	0.615	0.617	0.287	0.164	0.426	0.620	0.580	0.571	0.413	0.590
SPHINX-13B [18]	224^2	-	SFT	0.198	0.551	0.532	0.000	0.071	0.352	0.607	0.558	0.584	0.467	0.613
VisCoT-7B [35]	224^2	438k	SFT	0.355	0.610	0.719	0.279	0.341	0.356	0.671	0.580	0.616	**0.833**	0.682
VisCoT-7B [35]	336^2	438k	SFT	0.476	0.675	0.775	0.386	0.470	0.324	0.668	0.558	0.631	0.822	0.614
Qwen2.5VL-7B† [1]	336^2	-	-	0.350	0.642	0.735	0.202	0.472	0.325	0.603	0.556	0.455	0.347	0.616
Qwen2.5VL-7B† [1]	1536^2	-	-	0.773	0.710	0.792	0.492	0.708	0.663	0.685	0.604	0.457	0.371	0.603
Qwen2.5VL-7B* [1]	336^2	4k	SFT	0.355	0.658	0.740	0.215	0.489	0.334	0.624	0.563	0.467	0.405	0.619
Qwen2.5VL-7B* [1]	1536^2	4k	SFT	0.784	0.725	0.801	0.498	0.714	0.674	0.680	0.609	0.472	0.427	0.624
DocThinker-3B	336^2	4k	RL	0.460	0.663	0.746	0.213	0.486	0.335	0.664	0.572	0.486	0.485	0.625
DocThinker-3B	1536^2	4k	RL	0.751	0.691	0.762	0.469	0.735	0.566	0.682	0.583	0.490	0.517	0.637
DocThinker-7B	336^2	4k	RL	0.579	0.682	0.802	0.408	0.495	0.347	0.674	0.580	0.546	0.542	0.656
DocThinker-7B	1536^2	4k	RL	0.795	0.738	0.827	0.515	0.806	0.689	0.701	0.625	0.694	0.686	0.721
DocThinker-7B	1536^2	8k	RL	**0.802**	**0.757**	**0.836**	**0.568**	**0.814**	**0.697**	**0.734**	**0.641**	**0.737**	0.784	**0.768**

La valutazione avviene principalmente su due tipologie di compiti:

- **Compiti orientati al document undestranding**
- **Compiti di comprensione multimodale generale**

Si notano miglioramenti significativi in entrambi i casi, in particolare nel primo, e una forte capacità di generalizzazione.

4.3 Altri Risultati

Gli esperimenti mostrano risultati competitivi per ogni task, anche sui dati zero-shot dove supera tutti i modelli non basati su RL, mostrando grandi capacità di generalizzazione e adattabilità. Viene svolta una valutazione anche sul dataset TextREC che richiede la localizzazione di oggetti basandosi sulle richieste del testo.

Model	Template1	Template2
Specialist Models		
TransVG [6]	50.1	54.0
MAttNet [48]	52.3	60.5
QRNet [45]	52.7	59.1
MDETR [14]	54.4	63.3
TAMN [9]	77.8	80.8
DocThinker-7B	82.4	

4.3 Altri Risultati

I risultati mostrano come anche modelli specializzati come TAMN e MDETR vengono superati, rendendo chiaro che DocThinker eccelle nel ragionamento spaziale ed allinea efficacemente le informazioni testuali e visive, andando oltre il document understanding.

4.4 Studi di ablazione

Vengono svolti anche degli studi di ablazione sulle funzioni di ricompensa e sulla divergenza KL sui dataset DocVQA, TextCaps, TextVQA e InfoQA.
Negli studi sulle funzioni di ricompensa vengono eliminati:

- R_{RoI}
- $R_{rephrase}$
- Sia R_{RoI} che $R_{rephrase}$

Method	DocVQA	TextCaps	TextVQA	InfoQA
DocThinker-7B	0.795	0.738	0.827	0.689
w/o RoI IoU	0.775	0.693	0.803	0.637
w/o Rephrase Question	0.763	0.716	0.772	0.658
w/o RI & RQ	0.741	0.662	0.758	0.602

4.4 Studi di ablazione

- Eliminando R_{RoI} si nota un calo delle performance soprattutto in TextCaps e InfoQA evidenziando un ruolo importante nei compiti che richiedono un grounding visivo.
- Eliminando $R_{rephrase}$ il calo si nota principalmente su TextVQA e DocVQA espimendo un ruolo più importante nel chiarire le domande ambigue.
- Eliminando sia R_{RoI} che $R_{rephrase}$ si nota un degrado generale che sottolinea la crucialità delle ricompense multi-obiettivo per migliorare il ragionamento.

4.4 Studi di ablazione

Per lo studio dell'ablazione sulla divergenza KL si prova a porre semplicemente il coefficiente KL $\beta = 0$ poi $\beta = 0.001$.

Method	DocVQA	TextCaps	TextVQA	InfoQA
DocThinker-7B ($\beta = 0.04$)	0.795	0.738	0.827	0.689
w/o KL ($\beta = 0$)	0.780	0.719	0.803	0.676
$\beta = 0.001$	0.785	0.726	0.812	0.682

Con $\beta = 0$ si nota un calo generico delle prestazioni che evidenzia il ruolo fondamentale nel prevenire il catastrophic forgetting e rendere stabile l'addestramento.
con $\beta = 0.001$ si riscontra un leggero miglioramento rispetto al caso precedente ma senza raggiungere la stabilità iniziale.

Conclusioni

Il GRPO, grazie alle funzioni di ricompensa multi-obiettivo, supera le limitazioni del ragionamento statico basato su CoT e SFT.
Si conclude che DocThinker ottiene prestazioni migliori o altamente competitive sui benchmark standard rispetto ai precedenti modelli basati su SFT in molti compiti di document understanding.

References

- DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding - Wenwen Yu, Zhibo Yang, Yuliang Liu, Xiang Bai
- ChatGPT
- YouTube - Filippo Zanella - LLM vs MLLM: L'IA Capisce Davvero le Immagini? (Esperimento Reale!)
- Che cos'è la messa a punto? - IBM - https://www.ibm.com

Grazie per l'attenzione!