Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matěj Kripner		
Název práce	Self-Supervised Summarization via Reinforcement Learning		
Rok odevzdání	2024		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Milan Straka	**Role**	Oponent
Pracoviště	Institute of Formal and Applied Linguistics		

Text posudku:

The goal of the thesis is to explore reinforcement-learning methods for performing self-supervised summarization (i.e., not requiring reference summary). The main contributions of the thesis are the following:

- The author reimplements a pipeline for supervised finetuning and trains models of very good quality (slightly surpassing facebook/bart-large-xsum on XSum dataset; Chapter 3).
- The author implements a pipeline for reinforcement learning of pretrained generative models. The training uses a novel way of incorporating the KL divergence of the trained model and the pretrained one [Section 4.2.4], and the author proposes (to my best knowledge novel) token-level discounted rewards [Section 4.2.3]. A hyperparameter search is performed on a test task, arriving at hyperparameter values used in later experiments [Table 4.1].
- As the main theoretical contribution according to my view, the author proposes a novel method for computing dense reward function for summarization [Section 5]. Several variants are proposed (different kind of masking [Section 5.1.1] and weights of confusion coefficients [Sections 5.1.2-3]), and a detailed quantitative analysis is performed using two proposed metrics based on statistical testing [Section 5.2]. Results of a rich hyperparameter search is visualized in an innovative way [Figures 5.4, 5.5, A.1].
- Using all the above components, the author trains summarization models using reinforcement learning and the proposed dense reward function and evaluates them on XSum and CNN/Daily Mail datasets using 12 automatic metrics from SummEval [Section 6.2]. The models trained with reinforcement learning surpass the supervised-trained models on all reference-free metrics. Furthermore, the author perform manual evaluation [Section 6.3], showing that the RL-based models surpass the supervised-trained models and even the reference summaries in all three evaluated metrics (relevance, consistency, readability).

The thesis is written in very good English and is of outstanding quality. The performed experiments are on world-level and required vast amount of effort and also computational resources, considerably surpassing the required level for a Master thesis. Furthermore, the author proposed novel and successful theoretical approaches, demonstrating independent and high-quality scientific work in a world-wide relevant area.

I recommend the thesis to be defended.

Questions and Remarks

- Page 18, when using the 4 deviances as rewards, is the gradient computed through them, or is a detach() used? Both approaches would probably have a different training dynamic.
- Page 22, Figure 4.2, do you have an insight why the KL divergence increases very similarly to the reward (instead of staying for example roughly constant)?
- Page 39, Figure 6.1, while I understand that running BART-large on CNN/DM requires a lot of resources, why has it run for only 300 steps, compared to 800+ of the other configurations? Especially given that it has better results in Table 6.2 (compared to base RL), would you expect the results to improve if trained further? Relatedly, large RL on XSum does not seem to help; do you have an idea why (and possibly what effect the large-sized predictor has in CNN/DM and whether you think it would help on XSum too)?
- Page 40, shouldn't the second value in exponential_decay_length_penalty be larger than 1? That is how it is used in all examples; values < 1 should discourage generating an EOS.
- Page 41, Table 6.2, I am very much missing baselines for comparison – I found some in Table 3.2, but you should have definitely include existing models in the table.
- Page 41, Table 6.2, (Ouyang et al., 2022) use additional supervised loss to keep high scores on publicly available datasets even after RLHF (the "PPO-ptx" models); maybe a similar approach could be used to achieve both high reference-based and reference-free metrics.

Detailed Remarks to the Text

- Page 6, Equation 1.2, the on-policy distribution is only proportional to the right side, not equal.
- Page 6, Equation 1.3, I would consider using a_t for consistency, not just a.
- Page 14, Figure 3.1, it would be better not to plot values for 0K, allowing the range of the y axis to be more detailed (as in the top right figure).
- Page 15, Table 3.2 should mention also CNN/DM in the caption, not just XSum.
- Page 15, Table 3.2 would ideally include also models trained on 100% of the training data. You even have such models evaluated in Table 6.2 for XSum.
- Page 19, Equation 4.2, the $L_{\text{REINFORCE}}$ is not formally defined even it is said to be described in Section 1.2.1 – but there we maximize the objective, while here we need to minimize it (but I do not think it can cause confusion).
- Page 26, Figure 5.1, it would be nice to indicate word boundaries (e.g., with ##).
- Page 28, Figure 5.3, you say "All of these runs are stable and close to convergence." But for the large variants, it does not seem to me to be that close to convergence.
- Page 30, after equation 5.5, the $s_{\overline{X}1}$ should be $s_{\overline{X}t}$.
- Page 33, Figure 5.4, I really liked the figure. I think it could be further improved by
  - making the size differences to be larger
  - adding a legend with colors/shapes/sizes; currently the description is two pages away
- Page 33, Figure 5.5, you should indicate which results are base-sized and which large-sized.
- Page 38, BertScore is mentioned in the the text but not described at all.

Minor Remarks

- Page 8, APES is typeset in math mode with improper spacing.
- When typesetting quotes, ASCII quotes were sometimes used; instead, `` and '' should be used (p13 “ours/”, p18 “the”).
- Page 36, I would move Figures 5.6, 5.7, 5.9 one page earlier; furthermore, I would swap the order of Figures 5.8 and 5.9 (they are referenced 5.9 first and then 5.8).
- There are a few errors in commas in the text, mostly missing ones. For the sake of the author, here I list ones I wrote down, each described using a page and a small excerpt of text including the missing comma.
  p8 ROUGESal, which; p8 Entail, which; p9 Attention Based Credit, which; p13 Table 3.2, where; p16 main train method, which; p17 with our approach, which; p17 NumPy array, which; p17 [Bellman, 1957], where; p17 ends, when; p17 TRL and trlX libraries, in which; p18 R_n, the definition; p18 current batch, which; p19 [Abadi et al., 2016], we; p20 Llama 2 70B, but; p24 P(M_j), where; p25 -1), where; p25 never masked, since; p27 I^\omega, where; p28 hyperparameter configurations, we; p32 predictor size, where; p32 Welsh cyclist", which; p32 reward functions, which; p33 reward functions, we; p34 each summary, where; p34 of its creation, where; p38 this metric, we; p38 Fabbri et al., 2020], we; p41 SUPERT metric, which; p42 Repeated-3 metric, which; p42 generated summaries, but; p42 train data), and; p44 for each example, so; p44 Coherence metric, which; p45 was created, which; p45 first sentence, and therefore, the (both); p48 Section 5.3.2, where; p49 self-learning setting, where; p49 out-of-domain setting, where;

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 31. května 2024 **Podpis**