



دانشگاه تهران

دانشکده علوم و فنون نوین

گزارش هفته ششم و هفتم

نام و نام خانوادگی	فاطمه چیت ساز
شماره دانشجویی	830402092
تاریخ ارسال گزارش	6 بهمن 1402

Contents

1	برسی مقاله week to stronge learner
---------	------------------------------------

بررسی مقاله week to strong learner

در این هفته من سعی کردم یکم قشنگ تر مقاله رو بخونم

اولین سوالی که ذهنمو درگیر کرده اینه که اینا دقیقا چطوری میخوان align کنن و اینکه این دادن labelهای مدل ضعیف به قوی دقیقا چه میکنه

وقتی دوباره مقدمه رو خوندم دیدم خیلی تاکید کرده بود روی این قضیه که ما الان reinforcement learning داریم از human feedback و خب چون مدلا دارن میرن سمت superhuman دیگه شاید جواب نباشه خیلی اون یعنی اون مدلا از ما خیلی قوی تر هستند و خب ما میشیم یک سری supervisor که ضعیفه و خب ممکنه label هاش غلط باشه و حالا باید چه کرد ؟

یعنی این باعث شد که اون دید week to strong generalization بیاد

که ما ی سری مدل ضعیف داریم که میان label میدن در حالی که ما می‌خواهیم مدل قوی اونقدر تحت تاثیر قرار نگیره و وضعیتش نابود شه

مثلا مثال جالبی که خودش زده بود این بود که چار روز دیگه ی خروار کد برای ما میزنه اون مدل قویه و ما ب عنوان مدل ضعیف احتمالا نتونیم label درستی به مدل قویه بدیم

حالا برای این مسئله week to strong اومده مدل قوی رو با label های مدل ضعیف fine-tuning کرده

حالا چرا این کار میکنه ؟ چون الان مدل قویه میفهمه مشکلات مدل ضعیف چیه رو چه چیزایی حساسه و دقیقا حرفش چیه دیدش چیه و حتی ی جورایی ارورش هم میفهمه (به نظر من این عمق خطرناک بودنه)

ولی خب حرفشون اینه اون مدلی که اون تسک یا حالا کد سخت رو ساخته و درک اینو داشته احتمالا میدونه کدوم کد خوبه کدوم کد بد در واقع مدل ضعیف نیاز نیست بیاد فرایند یادگیری رو انجام بده فقط باید اون برانگیخته کنه اون چیزایی که همین الان مدل قوی می‌دونه

حالا تسک هایی که برای آزمایششون استفاده کردند

بیست و دو تا تسک nlp

تسک شطرنج

تسک reward model gpt هست

که برای مدل قویشون اومدن از gpt چهار استفاده کردند برای ضعیف gpt دو اما یک جا دیدم برای reward model اومده بود گفته بود ما مدل قویمون 3.5 عه که نفهمیدم چرا ؟

البته یک موضوعی که هست اینه خودشونم فهمیدن شاید الان فقط ی حال proof of concept دارن انجام میدن و توی قدم های اولیه هستند و ما نباید خیلی سخت بگیریم بهشون خلاصه که یک جورایی برداشت من اینه که preference learning اینا نیازمند ی label خفن از ادمه که چند وقت دیگه این زیاد شدنی نیست و خب باید حل کنیم این مشکل رو ی بحث جالبی که من زیاد نفهمیدم کرده بود scale oversight بود چیزی که فهمیدم این بود که اینجا برای حالاتی که ادم نمیتونه label خوب بده اومده از یک مدل دیگه خواسته نظرشو راجب خروجی مدل بده یا حتی مسئله رو به چندتا مسئله کوچک تر بشکنه یک همچین چیزی ولی دقیق نفهمیدم الان فرق این با مسئله ما چیه و چرا ما رفتیم سمت generalization و اصلا اینجا مفهوم generalization چیه بحث جالب بعدی eliciting latent knowledge هستند اینطوری که از مدل نظر صادقانه میپرسی (یعنی ک چی :/)

میگف ازش تو tampering اینا استفاده میکنن کلا این مفهوم صداقت و اینا ب نظر فان بود اما دقیق نفهمیدم عملیش چطور میشه موضوع بعدی بدبختی های قضیه بود اینکه الان ما هیچ دیدی نسبت به superhuman نداریم و این همه حرف میزنیم ممکنه الکی باشه بعد اینجا پریدم رفتم قسمت مدلس اینا که قضیش دقیقا چیه برای تسک nlp گفته بود بیست و دوتا تسک داریم همه چی رو هم باینری کرده و خب برای preprocessing|ش هم دیتا ها رو balance کرده و حتی ی سری ها رو حذف کرده برای مدل اومده قسمت un embedding رو حذف کرده جاش ی مدل خطی گذاشته که وزنش همون وزن un embedding هست و خروجیش دوتا چیزه که خب اینجا فکر کنم صفر و یک میشه من الان این un embedding رو زیاد نمیفهمم قضیش چیه

مدلشم طی دوتا ایپاک با بچ سائز سی و دو ترین میکنه و Early stop هم داره با توجه به label های week مدل دیتا ست هم نصف میکنه نصفشو آموزش میده به مدل ضعیف نصفشو تست میکنه به مدل ضعیف لیبل ها رو به عنوان soft label میده مدل قوی قضیع soft label رو من دقیق نفهمیدم دقیقا چیه نکته بعدی اینه نمودار سیزده صفحه سی و یک رو من نمیفهمم یعنی الان مدل ضعیف ها بهتر عمل میکنن؟ بعد این چطوری over fitting رو نشون میده

برای مسئله ی شطرنج اینطوریه که موقعیت مهره ها رو میدی میگ بهترین حرکتت چیه بعد اینا همه temperature صفره یعنی greedy محض دیگه اینکه پنجاهه هزار تا داده رو رندوم انتخاب برای week پنجاه هزار تا رو میدی week لیبیل میدی بعد میدی ب strong داده تست هم پنج هزار تا برای boostaping هم پنجاه هزار تا

قضیه اینه توی شکلش میاد نمایش میدی مال حالت zero shot رو و بعضی اوقات وقتی week خیلی ضعیفه عملکرد week to strong از zero shot هم کمتر میشه

و خب یک نکته ای که هست قضیه agreement هست

اینکه مدلامون (مدل student) وقتی بزرگ میشه agreement اش کم میشه

این قضیه agreement اینطوریه ک نگا میکنه مدل قوی چقدر داره از مدل ضعیف پیروی میکنه و خب منطقا اگ کپ هم عمل کنن pgr میشه صفر

اما خب ی نمودار باحال داره میگ وقتی مدل دانش آموز قوی میشه وقتی مدل معلم داره دری وری میگ زیاد agreement نمیکنه و اگ aux رو اضافه کنیم دیگ میشه نور علی نور و زیبایی مطلق

برای تسک بعدی یعنی reward model مربوط به gpt اینطوریه که ی دیتا داریم $(d, c1, c2, y)$ اینطوری که d میشه آخرین دیالوگ آمده از طرف یوزر و $c1$ و $c2$ میشه دوتا جوابی که مدل درست کرده و y اگر یک باشه یعنی ترجیح $c2$ صفر باشه یعنی یکی دیگ

حالا اینو برای هر دوتا $c1$ و $c2$ انجام میدن مدلش اینطوریه باز اومده لایه un embedding رو دست کاری کرده به خطی با یک خروجی میگ این خروجیه logit عه

بعد برای هر دو ی بار مدلو فوروارد و خروجی میشه اختلاف سیگموئید این دوتا logit ها: F.Ch > قضیه بعدی generative fine-tuning هست که ما بیایم یک سری چیز رو برجسته کنیم برای مدل ها بدون اینکه بهشون label بدیم در واقع یک فرایند unsupervised هست به صورت fine-tuning با استفاده از تسک های مرتبط به تسک اصلی

مثلا خودش مثالی که زده بود این بود ک ما ی مدل زبانی داریم حالا میخوایم به احساسات اهمیت بیشتری بده پس بهش مثلا کامنت های آنلاین ملت رو میدیم ک دستش بیاد احساسات چطوریه اما من اصلا نمیفهمم چطوری؟

حالا برای همین مسئله reward model ما اینطوریه ک ما مدل قوی و ضعیف رو هر دو بعد training میایم با generative fine-tuning آموزش میدیم ک زیبا میکنه انگار مدل ها رو

ساز و کارشم اینطوریه ک مثلا یک پیام یوزر و دوتا جواب براش رو می دیم و خب انگار این دوتا جواب بدترین جواب های ممکن بوده ولی ترجیح یوزر رو بهش نمیدیم

قضیه اینه الان مدل یاد میگیره چیو///:

بدترین جواب رو؟

ی حرف دیگه هم زده بود اینکه ممکنه leakage داشته باشیم اینم نمیفهمم چرا

البته این مفهوم leakage رو میاد یکم بعدتر توضیح میده که مثلا سوالاتی ما ممکنه تو داده های

pretrain gpt باشه و خب این خوب نی باعث میشه نتایج ما الکی خوب باشه

اما ی حرف عجیبی میزنه اونم اینه که برای مدل هلی superhuman اینطور نیست اونا اصل داده رو

ندارند حالا ممکنه ی چی شبیهشو داشته باشه و تقلیدی وجود نداره

البته گفته ما برای proof of concept اومدیم ی مدلی رو نشون دادیم که داده train اش فرق داشته و از

این حرفا

اما میگم جالبیش این بود که میگف ما در superhuman ها وضعمون اینطور نیست ما اونجا self

supervise هستیم

ی حرف دیگشم این بود که prompting اگ استفاده کنیم ممکنه بیشتر تحت تاثیر این قضیه قراره

بگیره

حالا سوال پیش میاد prompting چیه

اینطوری که مثالی حال سوالی طور از مدل میپرسه که نظرش راجب فلان چیز چیه بعد جواب مدل رو

برسی میکنه دیگ اون دردرسهای label بده بره و بیاد رو نداره

یا حتی میشه few shot قضیه رو دید البته

اما ما چون زیاد اعتماد ب دیتا pretrain مون نیست این زیاد به دردمون نمیخوره فعلا

Dataset	Custom Prompt
CosmosQA	There is some context, followed by a question about the context, and an answer. Respond with 1 if the answer is correct, and 0 otherwise.
Deontology	There is a short scenario, consisting of a situation or request, followed by an excuse. Respond with 1 if the excuse is valid, and with 0 otherwise.
Justice	There is a statement. Respond with 1 if it is correct or just and with 0 otherwise.
QuAIL	There is a news context, followed by a question about the context, and an answer to the question. Respond with 1 if the answer is correct, and with 0 otherwise.
SciQ	There is a science knowledge question, followed by an answer. Respond with 1 if the answer is correct, and with 0 otherwise.
Social IQa	There is some context, followed by a social question, followed by an answer. Respond with 1 if the answer is correct, and 0 otherwise.
Virtue	There is a short scenario, followed by a judgement of the person involved. Respond with 1 if the judgement is correct, otherwise respond with 0.

خب حالا یکم راجب auxiliary confidence صحبت کنیم

فرمولش ک بدین صورته

$$L_{\text{conf}}(f) = (1 - \alpha) \cdot \text{CE}(f(x), f_w(x)) + \alpha \cdot \text{CE}(f(x), \hat{f}_t(x))$$

قضیه اینه ما ی cross entropy داریم ک میاد توزیع بین مدل ضعیف و مدل قوی رو برسی میکنه و این میشه loss اش حالا بحث بعدی اینه ما برای اینکه قضیه رو متعادل تر کنیم ی hardened strong model prediction میسازیم ک خروجیش یا یکه یا صفر و اینطوریه ک اگ $f(x)$ بیشتر از یک threshold باشه یک وگرنه صفر برمیگردونه و خب میگه این کمک میکنه به balance قضیه در قسمت بعد اومده ی سری روش هایی ک خیلی جواب نداده اما شاید جواب بده بعدا رو معرفی میکنه اولیش اینه ی threshold بدیم بگیم اینا رو مدل ضعیف خوب فهمیده حالا بیا با همین اینایی ک خوب فهمیده مدل قوی رو آموزش بده ولی این مال وقتی که مدل ضعیف خیلی وضعیت داغونی داشته باشه و وقتی مدل ضعیف بدک نیست اونقد تاثیری نداره

روش بعدی productive confidence loss هست که توش میخوایم cross entropy ما له حاصل ضرب دوتا احتمال های مدل ضعیف و قوی برسه که خب اینم بعضی جاها خوب بود بعضی جاها زیاد خوب نبود

موضوع بعدی اینه که بیایم یک مدل خطی از مدل قوی بسازیم بعد تمام لایه ها رو با مقادیر اون fine-tuning کنیم

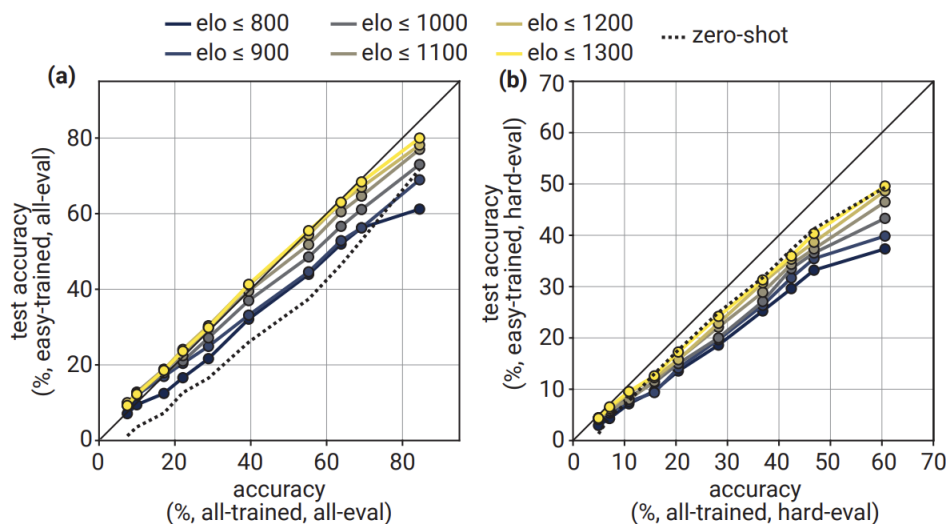
روش های بعدی weight regularization هست مثل lora اینطوری که مثلا یک مدل خطی روی پارامتر ها داریم و ی جورایی همه وزن ها رو نگهداری نمیکنیم انگار به صورت محدودتری میایم fine-tuning میکنیم

روش های data augmentation اینطوریه که مثلا یک مدل قوی داریم که میاد داده ها رو تقویت میکنه بعد خروجی این مدل قوی و مدل اصلی باهم مقایسه میشه و خب axillary loss رو تفاوت این دوتا ایجاد میکنه

مسئله بعدی easy to hard هست که جالبه مثلا اگر ما مسئله رو با سوالای ساده برای مدل ضعیف fine-tuning کنیم عملکردش از zero shot بهتره

اما برای مسائل سخت اگر این کارو بکنیم اونقد کارمون با zero shot فرقی نداره

در واقع جالبه وقتی مدل در سوالاتی سخت حتی هر چی مدل ضعیف یادگرفته هم میگیره بازم اونقد از zero shot بهتر نی



مسئله بعدی اینکه ما داریم دقیقا چی کار میکنیم:

یعنی الان چطوری میتونیم یک موضوعی را بررسی کنیم وقتی مشکل رو الان نداریم

خب باید ی سری چیز رو رعایت کنیم

مثلا فرضیاتمون مشخص باشه و سعی کنیم برای فرض های مختلف راه حل ارائه بدیم

بدونیم برای هر چیز چه گیر و گور هایی داریم

راه حل هامون scalable باشه

در واقع یک دلیلی هم ک رفتن سراغ fine-tuning نرفتن سراغ few shot همینه

توی fine tuning اینطوریه اونی که مدل الان میدونه رو میخوایم دقتش رو ببریم بالا بدون اینکه بدونیم

اینی که میدونه رو از کجا یاد گرفته

اما تو few shot اینطوری نی میگ با اطلاعات کم بیا یک توزیع گسترده از داده ها رو بساز

حالا بعدش یکم راجب این میگ ک مسیر چیه

یک اینکه ما باید بیایم align کنیم

حالا راه حل های موجود

RLHF

Scalable oversight

Constitution ai

Adversarial training

ک ما اومدیم گیر دادیم به RLHF اینجا

بعد حالا که align کردیم بررسی کنیم آیا درست align کردیم یا نه ؟

روش های موجود

Red teaming

Interpretability

بعد همین فرایند رو هی انجام بدیم و بزرگ و بزرگ تر کنیم با bootstrapping

ی سری چیزا هم برای ما مهمه مثل

امنیت

صداقت

رعایت پروتوکل های انسانی

امنیت کد

که توقع میره که اون مدل های قوی خیلی درک بالایی از اینا داشته باشند

در واقع به صورت ایده ال اونا انقد این مفاهیم رو میفهمند که میشه تبدیلیش کرد به یک reward model

ولی اگر اینطور نشد میشه از یک مفهومی به نام oracle استفاده کرد

در واقع حرفشون بوی این رو میداد که اگر انقدر superhuman چرا از خودش نپرسیم که وضعیت

چطوره و چطوری میشه مشکل را حل کرد