Conclusie Seminarie Turing Test

Sil Vaes

Contents

1 Is de Turing Test nog relevant? **1**
1.1 Task-oriented evaluation 2

2 Notes **4**

1 Is de Turing Test nog relevant?

Niet echt, eerder een gedachtenexperiment. Ook raar dat hij een positivist was, maar de regels om te slagen zijn open voor interpretatie. Ook was schaken ook een van de intelligentietests, maar naar Deep Blue is dit geschrapt. Test niet echt intelligentie maar "menselijkheid". Er worden oppervlakkige conclusies getrokken. "Een computer kan gediagnosticeerd worden met koorts omdat de CPU een hoge temperatuur heeft". Ook raar dat een persoon met zijn overtuigingen, logisch positivisme, geen objectieve metriek heeft bedacht om de machine te beoordelen. Misschien een reden dat dit essay/paper gepubliceerd is in een blad voor filosofie.

Wat zijn dan wel moderne en relevante technieken om AI te evalueren? Twee soorten aanpakken:

- Task-oriented: met drie soorten evaluatie: human discrimination, problem benchmarks en peer confrontation.
- Ability-oriented: Waar een systeem beoordeeld word op zijn (cognitieve) vaardigheden ipv de taken die het kan vervolledigen. "the adaptation of cognitive tests used for humans and animals, the development of tests derived from algorithmic information theory or more integrated approaches under the perspective of universal psychometrics."

Waarom twee soorten aanpakken? Er zijn twee clashende definities van AI:

- "[AI is] the science and engineering of making intelligent machines" - McCarthy.
- "[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]" - Minsky.

De eerste is duidelijk ability-oriented en de tweede task-oriented. Iets ability-oriented evalueren is ook beduidend moeilijker. Bij task-oriented moet men gewoon kijken of de taak correct wordt uitgevoerd, lijkt ook meer op de Turing Test. Een AI voert een gespecialiseerde taak dan ook zonder intelligentie uit, wat een paradox eigenlijk. Dit is ook een reden van het succes, het is voorspelbaar. Zoals we in het begin van de cursus hebben gezien: "Iets is een AI probleem totdat het is opgelost". Dit wordt ook wel het "AI effect" genoemd (McCorduck). Voor dit seminarie is enkel het verschil tussen narrow en general AI belangrijk. Narrow of specialised AI heeft task-oriented evaluation nodig en general-purpose AI heeft ability-oriented evaluation nodig. Maar sommige AGI worden op specifieke taken getest en narrow AI's worden met general-purpose componenten zoals planning/learning geintegreerd en worden na trainingscycli gespecialiseerd en verliezen zo plasticiteit.

Zoals in alle wetenschap zijn metingen cruciaal. Momenteel zijn er benchmarks en competities, maar er is nog ruimte voor groei. Deels omdat men AI evaluatie niet echt als een meting ziet. Ook is het belangrijk om van de task-oriented mentaliteit af te geraken omdat er veel vooruitgang is in meer general AI met de opkomst van developmental robotics, deep learning, inductive programming, AGI, etc.

1.1 Task-oriented evaluation

Algemene principes worden toegepast, maar de taak is specifiek genoeg om een gespecialiseerd systeem te maken. Van een abstract probleem naar een specifieke taak wordt aangemoedigd. Wanneer we een specifieke taak hebben, moeten we een notie van performance hebben. Deze wordt dan ook gemeten. Het is duidelijk dat hier niet echt intelligentie wordt gemeten. Een AI voltooid deze taken ook op een heel andere manier dan mensen. We zeggen dan ook bijvoorbeeld dat het team achter Deep Blue Kasparov heeft verslaan.

Performance van een systeem wordt gemeten in een white-box approach als de oplossing relatief eenvoudig is. Een typisch probleem waarbij dit gedaan wordt is wanneer de oplossing correct of optimaal moet zijn (perfect dus). Dus eerder klassieke computational complexity theory.

Maar tegenwoordig zijn AI oplossingen zo complex dat perfecte oplossingen niet meer een constraint zijn. Dus zijn er approximate solvers die een performance metriek optimaliseren. Hier is average case analyse ook relevanter, maar bij sommige paradigma's wordt worst case ook bekeken (Probably Approximately Correct learning). White-box evaluation wordt ook vaak gebruikt bij spellen te spelen.

Maar tegenwoordig is AI zo complex geworden dat de white-box approach niet meer haalbaar is. Black-box approaches kunnen in drie categorien ondergebracht worden:

- Human discrimination: het AI wordt getest door observers. Komt niet vaak voor, hier valt de Turing Test onder. (Komt wel voor bij psychologie)
- Problem benchmarks: het systeem wordt getest op een repository van problemen. Dit komt vaak voor, er zijn libraries en bestaande repositories voor. Deze zijn ook vaak public, maar kan leiden tot "evaluation overfitting" (gespecialiseerd op deze problem-set).
- Peer confrontation: het systeem wordt getest in een serie van n-vs-n matches. Komt ook vaak voor, performance is relatief tov andere systemen. Vooral bij spellen en multi-agent systems.

Een combinatie van de drie kan ook. Common-sense reasoning kan door alle drie geevalueerd worden bijvoorbeeld. (interviewing, comprehension tests en een game show bijvoorbeeld).

Human Discrimination:

De Turing Test was bedoeld als een filosofisch instrument om tegen de contra argument van machine intelligence tegen te gaan, dus nooit bedoelt als een echte test. Het is een test die de menselijkheid test, niet intelligentie. Dit betekent niet dat we menselijke oordeel uit testen moeten halen. Er zijn varianten die wel nuttig zijn. (Total Turing tests (Schweizer 1998), Visual Turing tests including sensory information, Toddler Turing tests (Alvarado et al. 2002), robotic interfaces, virtual worlds, etc. (Mueller and Minnery 2008; Hingston 2010).) Deze zijn nuttig om chatbots, personal assistant, videogames etc te testen. Als in: zijn ze geloofwaardig?

Problem benchmarks:

Als de problemset klein en/of bekend is kan er een grote switch in het systeem gezet worden. Wanneer het systeem het probleem herkent kan het alsnog een hardwired oplossing gebruiken. -> Evaluation overfitting. Nog een groter probleem als de problemset gemaakt is door de researchers zelf.

Bijvoorbeeld een self-driving car ontwikkelen op een kleine parking. Beter: een grote en diverse problemset. Nog beter: infinite, dus een generator. Maar hoe moeten we een infinite problemset evalueren? Eerste n problemen evalueren? Random sampling lijkt beter. Twee varianten:

- Information-driven sampling: Via clustering zo een divers mogelijk aantal problemen samplen via similarity.
- Difficulty-driven sampling: Moeilijke problemen samplen (via tijd nodig of error).

Beide manieren kunne adaptief gemaakt worden.

Peer confrontation:

We evalueren een systeem door het te laten concureren met een ander systeem. Er wordt een match gespeeld tussen meerdere systemen. Dit komt vaak voor in spellen en bij multi-agent research. Het resultaat, van soms meerdere metingen, wordt gebruikt om te bepalen welk systeem het beste is. Zoals duidelijk is, is het grootste probleem dat alle metingen relatief zijn. Ondanks dit relatieve karakter kan men nog steeds een gemiddelde performance bepalen. Het grootste probleem is robustheid en standaardisatie van de resultaten. Hoe bepaalt men de resultaten van twee "competities" als de deelnemers anders zijn? Gebruik rankings zoals Elo of meer geavanceerde systemen. Het beste is ook om competitie na competitie voorgaande tegenstanders terug mee te laten doen. Let wel op dat systemen niet gaan specialiseren om tegen bepaalde tegenstanders te spelen.

Een alternatief is om systemen tegen gestandaardiseerde tegenstanders te laten spelen. Hoe kiezen en hoe ook weer hetzelfde specialisatieprobleem tegen gaan? Voorgaande spelers met andere parameters? Tegenstanders laten cheaten met meer info?

Highlights en toekomst:

2 Notes

Wel: "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement" -> *We identify three kinds of evaluation: Human discrimination, problem benchmarks and peer confrontation.* [1] `https://sci-hub.mksa.top/10.1007/s10462-016-9505-7`

"various video game platforms for AI evaluation" -> `https://ojs.aaai.org//index.php/aimagazine/article/view/2748`

Measuring progress: `https://www.eff.org/ai/metrics`

References

[1] José Hernández-Orallo. "Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement". In: *Artificial Intelligence Review* 48.3 (2016), pp. 397–447. DOI: 10.1007/s10462-016-9505-7 (cit. on p. 4).