

Vývoj toolkitu pro red-teaming velkých jazykových modelů (LLM)

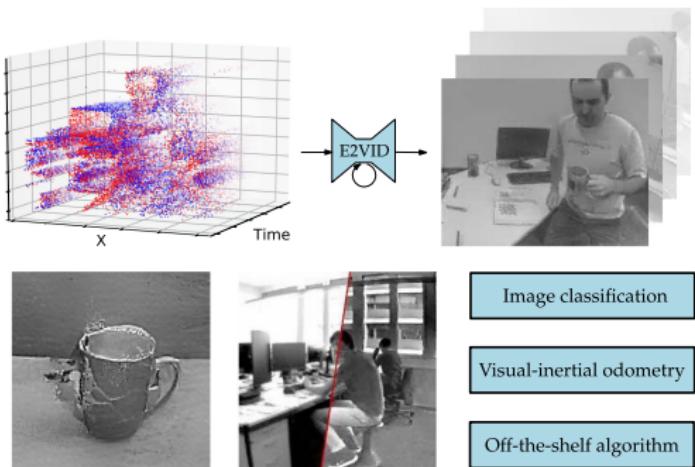
Adam Veselý

Vedoucí: Ing. Jakub Reš

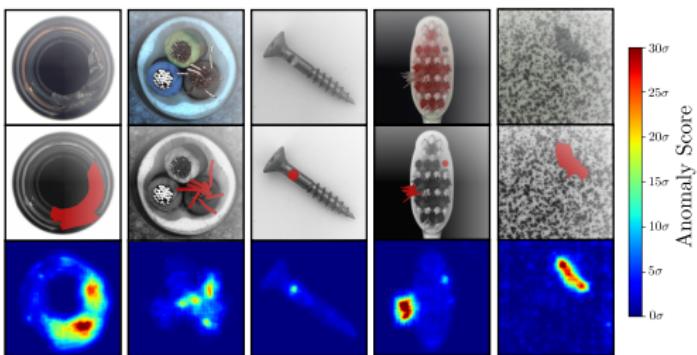


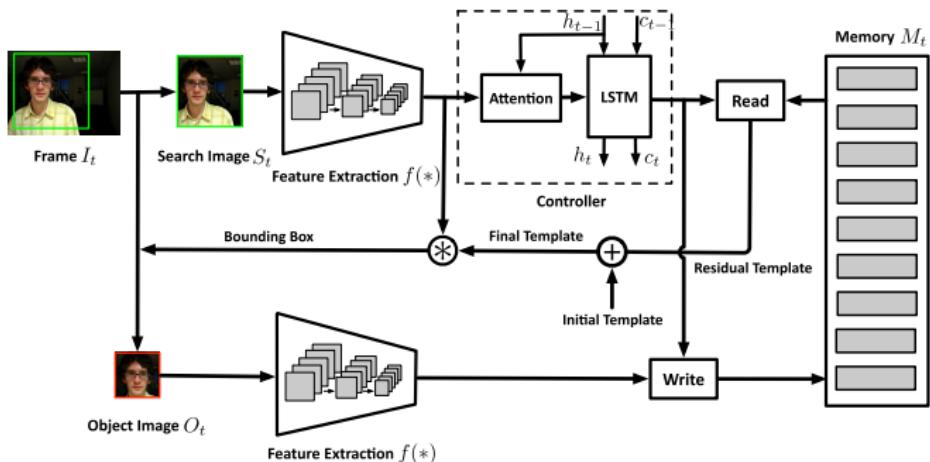
17. ledna 2026

- Vstupy či stav před
- Co mají být výstupy
- Odrážky žádné nebo aspoň stručné!
- Žádoucí: Schéma se vstupy a výstupy



- Vstup
- Výstup
- Žádoucí vlastnosti
- Využití & aplikace





$$\mathbf{a}_t = \sum_{i=1}^L \alpha_{t,i} \mathbf{f}_{t,i}^* \quad (1)$$

kde $\alpha_{t,i}$ počítá softmax:

$$\alpha_{t,i} = \frac{\exp(r_{t,i})}{\sum_{k=1}^L \exp(r_{t,k})} \quad (2)$$

$$r_{t,i} = W^a \tanh \left(W^h \mathbf{h}_{t-1} + W^f \mathbf{f}_{t,i}^* + b \right) \quad (3)$$

Podstatné informace o řešení



Sablonu prezentace ZP - Online X + https://www.overleaf.com/project/

Sablonu prezentace ZP

Source Rich Text

File outline

We can't find any sections or subsections in this file. Find out more about the file outline

trochu přiblížte, aby bylo zřejmo, o co jde, ale nevysvětlujte je podrobněji. Měli byste posluchači rozuměli a dokázali ho naprogramovat, když byste jim vysvětlili, na čem pracujete a jak se vám to dělá.

% -- Podrobnosti návrhu vašeho systému. Opět, posluchači nebudu vás systém hackovat, nepotřebují detailní strukturu tříd, názvy funkcí, jména souborů, datové formáty apod. Tyto věci uvádějte pouze v takové míře, která pomůže posluchačům učít si představu, na čem pracujete a jak se vám to dělá.

% HEROURL, Adam. Prezentování. *Herout.net: Poznánky učitele, kouče, čtenáře [online]*. [cit. 2021-9-15]. dostupné z: <https://www.herout.net/blog/category/prezentovani/>

%

% Uveďte, jaké zajímavé problémy jste v práci řešili.

% Mělo by z tohoto být patrné, že je to závěrečná práce -- ne jen další projekt do předmětu -- tedy že v tomto něco netriviálního, zajímavého a přínosného.

% Raději dva nebo tři slajdy, které ukážete/vysvětlíte během 20-vteřin, než se snáší všechno namastit na jednu slajdu.

% Na slajdy je dobré dat vizuální informaci: vzorce, schéma, obrázky, diagramy. Slovní informaci můžete předat peskem.

Slovní informaci můžete předat peskem: vzorce, schéma, obrázky, diagramy. Slajdy můžete využít k vysvětlení principu, kterým je řešení.

\begin{frame}\frametitle{Podstatné informace o řešení}

\centering\includegraphics[width=0.8\textwidth]{img/template-Schema.pdf}

\begin{equation}

\mathbf{f}_t = \sum_{i=1}^L \alpha_i \mathbf{f}_{t,i}

kde α_i počítá softmax:

\alpha_i = \frac{\exp(r_{t,i})}{\sum_{k=1}^L \exp(r_{t,k})}

\mathbf{r}_t = W^0 \tanh((W^0 \mathbf{h}_{t-1} + W^1 \mathbf{f}_{t-1} + b))

Celý výběr poskytovaný pomocí předních vrstev je deštěr, aby se nedostal do fronty.

Recompile

Review Share Submit History Chat

Podstatné informace o řešení

Podstatné informace o řešení

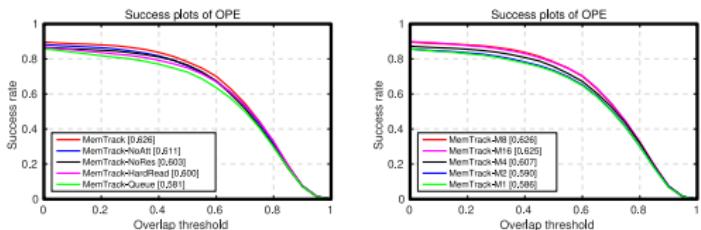
Podstatné informace o řešení

Vývoj toolkitu pro red-teaming velkých jazykových modelů (LLM)

5/7

- Co se podařilo
- Vytvořená datová sada: **105 k** záznamů
- Úspěšnost: **103 %**

	AN	RN	EAO ↑	$\mathcal{J}_M \uparrow$	$\mathcal{F}_M \uparrow$	Speed
SiamFC	✓		0.188	-	-	86
SiamFC		✓	0.251	-	-	40
SiamRPN	✓		0.243	-	-	200
SiamRPN		✓	0.359	-	-	76
SiamMask-2B w/o R	✓		0.326	62.3	55.6	43
SiamMask w/o R	✓		0.375	68.6	57.8	58
SiamMask-2B-score	✓		0.265	-	-	40
SiamMask-box	✓		0.363	-	-	76
SiamMask-2B	✓		0.334	67.4	63.5	60
SiamMask	✓		0.380	71.7	67.8	55



Sabíra prezentaci ZP - Online

https://www.overleaf.com/project/

Source RichText

Sabíra prezentaci ZP

File outline

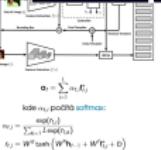
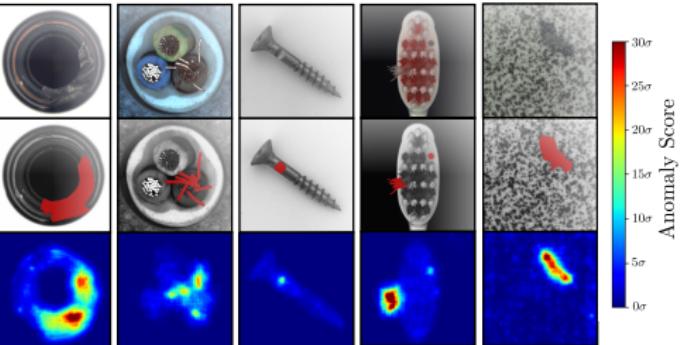
- String
- image_graphic.png
- example_cz.pdf
- filolog1_cz.pdf
- filolog1.pdf
- filolog2.pdf
- placeholder_1.jpg
- placeholder_2.jpg
- placeholder_3.jpg
- questions_cz.pdf
- questions_en.pdf
- security.jpg
- smile.jpg
- template-Goal.pdf
- template-Screens.pdf
- File outline

We can't find any sections or subsections in this file.
Find out more about the file outline.

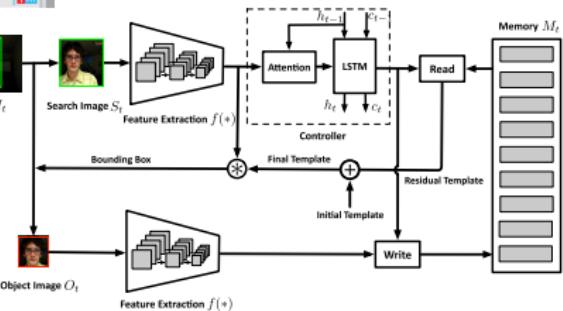
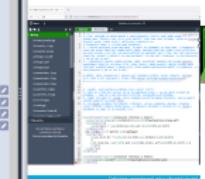
```

1 Drážděte až dole až nejdřív, a co jde, ažte nevysvěcujete zejména
2 drážděním, když test posloužíte algoritmu, respektive k detekci, že něco je
3 výjimečné. Podrobností návrhu vašeho systému, dleto, počítáváte několik výkazů
4 až výkazů, které vám poskytují informace o tom, co se dělá. Nejdřív, dokud
5 nejdřív, ažto vás vede k výkazu, který má všechny součásti, dostatečně
6 určitě a přesné, na čem pracujete a jist se vše v tom dořeší.
7 NEDĚLÍ: Až do prezentace, vydáte své: Pouze aktuálně, když, čtená.
8 [Záloha]: [Téma: 2023-8-12], vydávajíte:
9 https://www.hradat-metodika.cz/seznam/referencni/
10
11 R - Lépe, zdejší zadání projekty jste v práci řešili,
12 až do této chvíle, když jste všechny výkazy vytvořili – ne jen další projekt do
13 představení – tedy je v tom náročně až extrémně, zájmivou a pohromadě.
14 R - když dnes už máte všechny výkazy třeba 20+ výkazů, ned je
15 výkaz, který je dobré dát vizuální referenci; výkazy, schématy, obrázky, diagramy,
16 životního cyklu, mohou přinést pomoc. Je dobré zahrát a odrážet mnoho různých výkazů v
17 souboru, než vše, co vystavíte ruce.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

```



Podstatné informace o řešení



- Pokud je otázek více, lze udělat i více slajdů.
- Tento slajd nechť je příloha, která se nepočítá do celkového počtu slajdů.
- Otázku je dobré sem přepsat **verbatim**, ať není pochybnost, jestli nedošlo k nepřesnému parafrázování.

