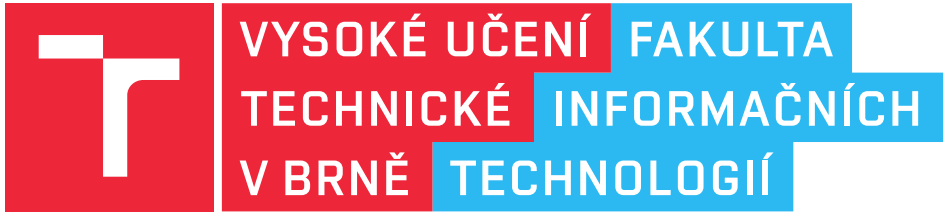
Vývoj toolkitu pro red-teaming velkých jazykových modelů (LLM)

Adam Veselý

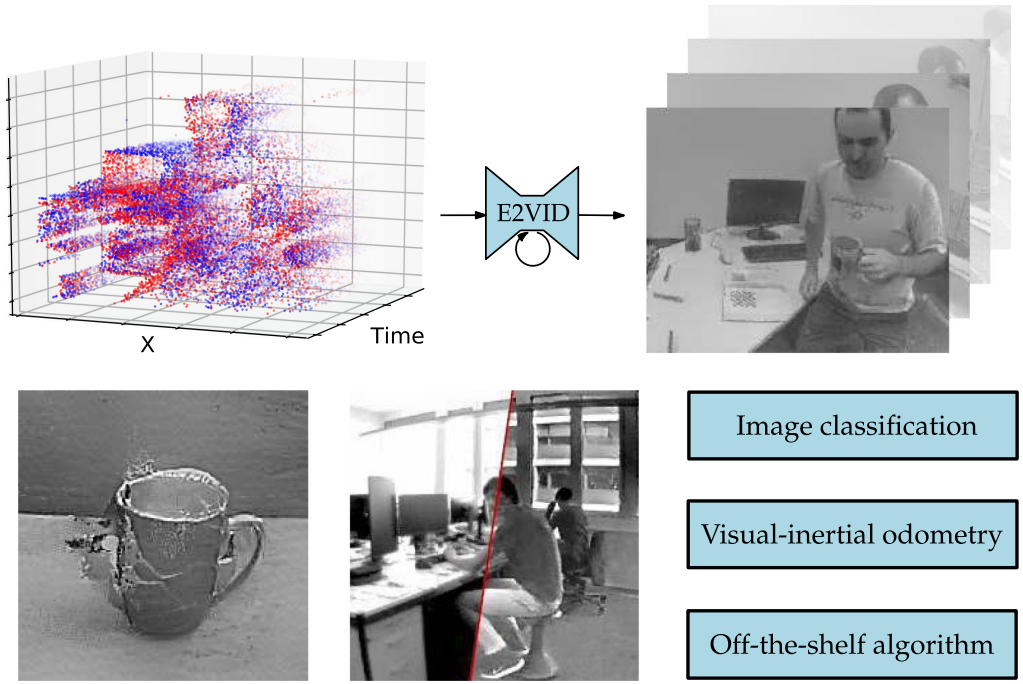
Vedoucí: Ing. Jakub Reš



17. ledna 2026

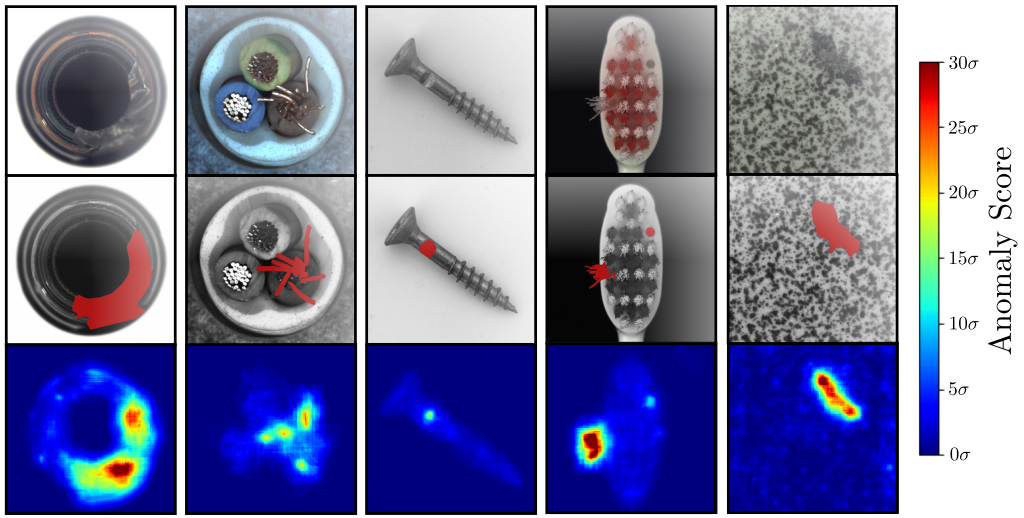
Motivace

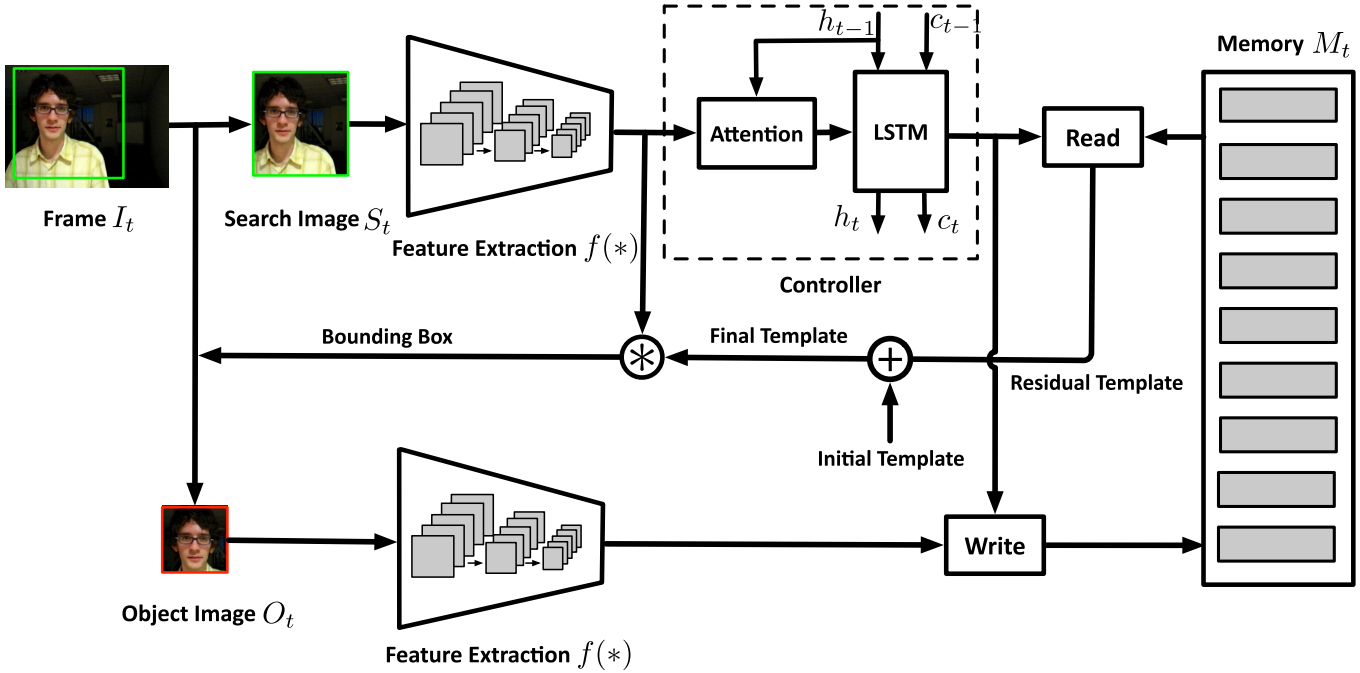
- Vstupy či stav před
- Co mají být výstupy
- Odrážky žádné nebo aspoň stručné!
- Žádoucí: Schéma se vstupy a výstupy



Cíle práce

- Vstup
- Výstup
- Žádoucí vlastnosti
- Využití & aplikace





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

kde $\alpha_{t,i}$ počítá softmax:

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

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

Podstatné informace o řešení

Sablona prezentace ZP - Online

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

Menu | Sablona prezentace ZP | Review | Share | Submit | History | Chat

Source | Rich Text | Recompile

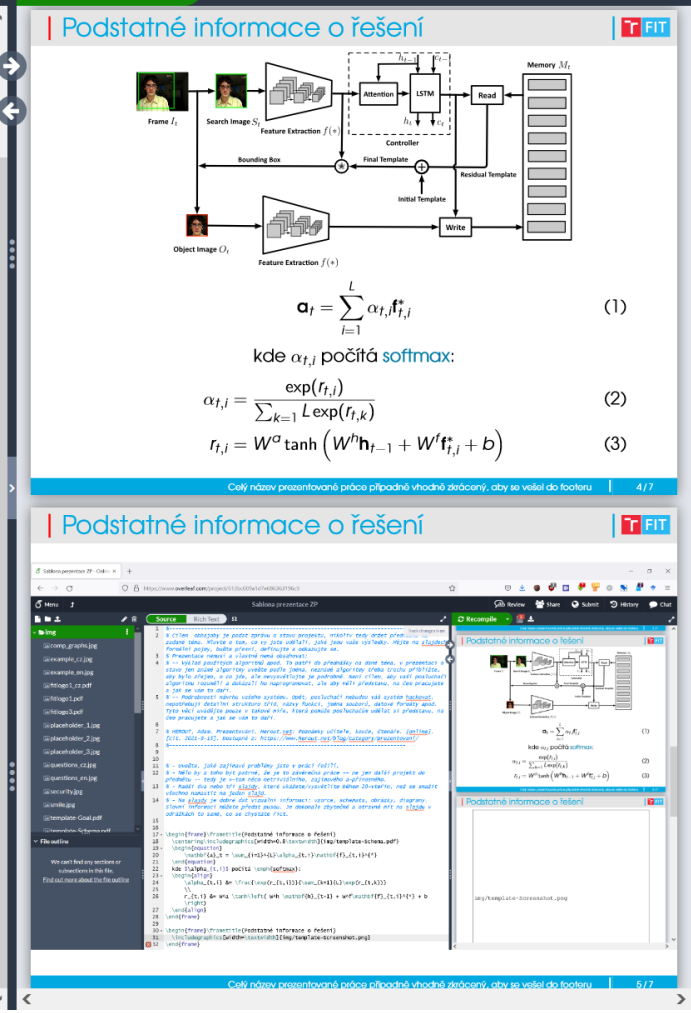
Track changes is on

```
trochu přiblížte, aby bylo zřejmo, o co jde, ale nevysvětlujte je podrobně. Hlavní
cílem, aby vaši posluchači algoritmu rozuměli a dokázali ho naprogramovat, ale aby
měli představu, na čem pracujete a jak se vám to daří.
% -- Podrobnosti návrhu vašeho systému. Opět, posluchači nebudou váš systém
hackovat, nepotřebují detailní strukturu tříd, názvy funkcí, jména souborů, datové
formáty apod. Tyto věci uvádějte pouze v takové míře, která pomůže posluchačům
udělat si představu, na čem pracujete a jak se vám to daří.

% HEROUT, Adam. Prezentování. Herout.net: Poznámky 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 toho být patrné, že je to závěrečná práce -- ne jen další projekt do
předmětu -- tedy je v~tom něco netriviálního, zajímavého a~přínosného.
% - Radší dva nebo tři slajdy, které ukážete/vysvětlíte během 20~vteřin, než se
snažit všechno namastit na jeden slajd.
% - Na slajdy je dobré dát vizuální informaci: vzorce, schemata, obrázky, diagramy.
Slovní informaci můžete předat pusou. Je dokonale zbytečné a otravné mít na slajdu v
odrážkách to samé, co se chystáte říct.


\begin{frame}\frametitle{Podstatné informace o řešení}
  \centering\includegraphics[width=0.8\textwidth]{img/template-Schema.pdf}
  \begin{equation}
      \mathbf{a}_t = \sum_{i=1}^{L}\alpha_{t,i}\mathbf{f}_{t,i}^{*}
  \end{equation}
  kde $\alpha_{t,i}$ počítá \emph{softmax}:
  \begin{align}
      \alpha_{t,i} &= \frac{\exp(r_{t,i})}{\sum_{k=1}^{L}\exp(r_{t,k})}
      \\
      r_{t,i} &= W^a \tanh\left( W^h \mathbf{h}_{t-1} + W^f\mathbf{f}_{t,i}^{*} + b
      \right)
  \end{align}
\end{frame}

\begin{frame}\frametitle{Podstatné informace o řešení}
  \makebox[\linewidth]{\includegraphics[width=\paperwidth]{img/template-Screenshot.p
  ng}}
\end{frame}
```

img
- comp_graphs.jpg
- example_cz.jpg
- example_en.jpg
- fitlogo1_cz.pdf
- fitlogo1.pdf
- fitlogo3.pdf
- placeholder_1.jpg
- placeholder_2.jpg
- placeholder_3.jpg
- questions_cz.jpg
- questions_en.jpg
- security.jpg
- smile.jpg
- template-Goal.pdf
- template-Schema.pdf

File outline

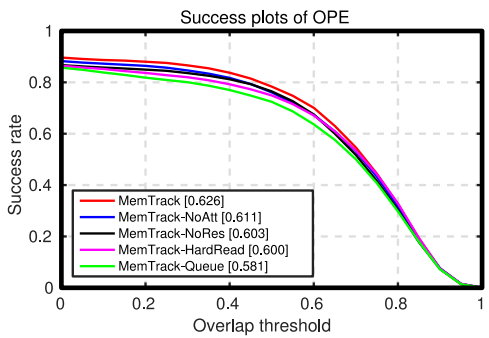
We can't find any sections or subsections in this file.
Find out more about the file outline

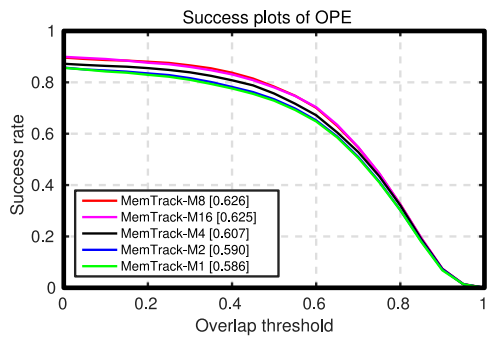


Výsledky práce

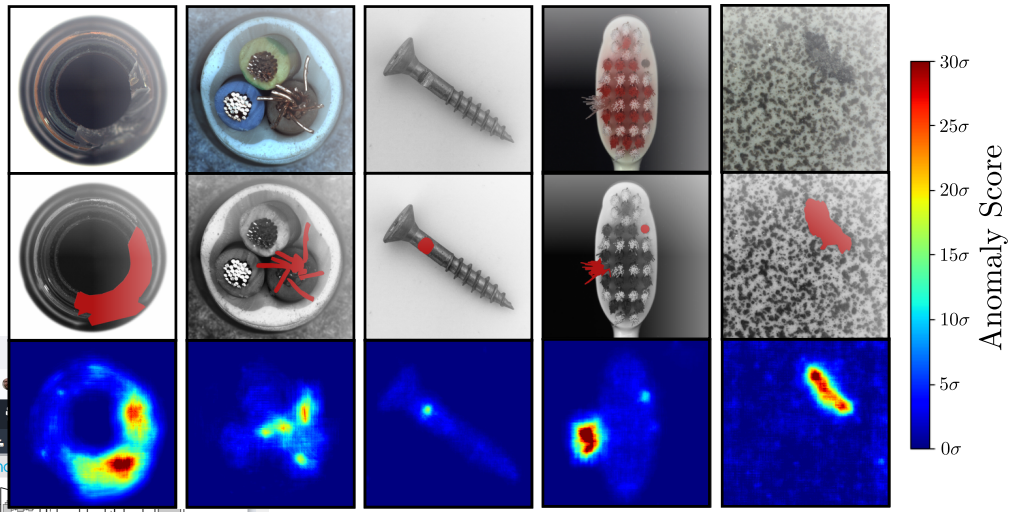
- Co se podařilo
- Vytvořená datová sada: 105 k záznamů
- Úspěšnost: 103 %

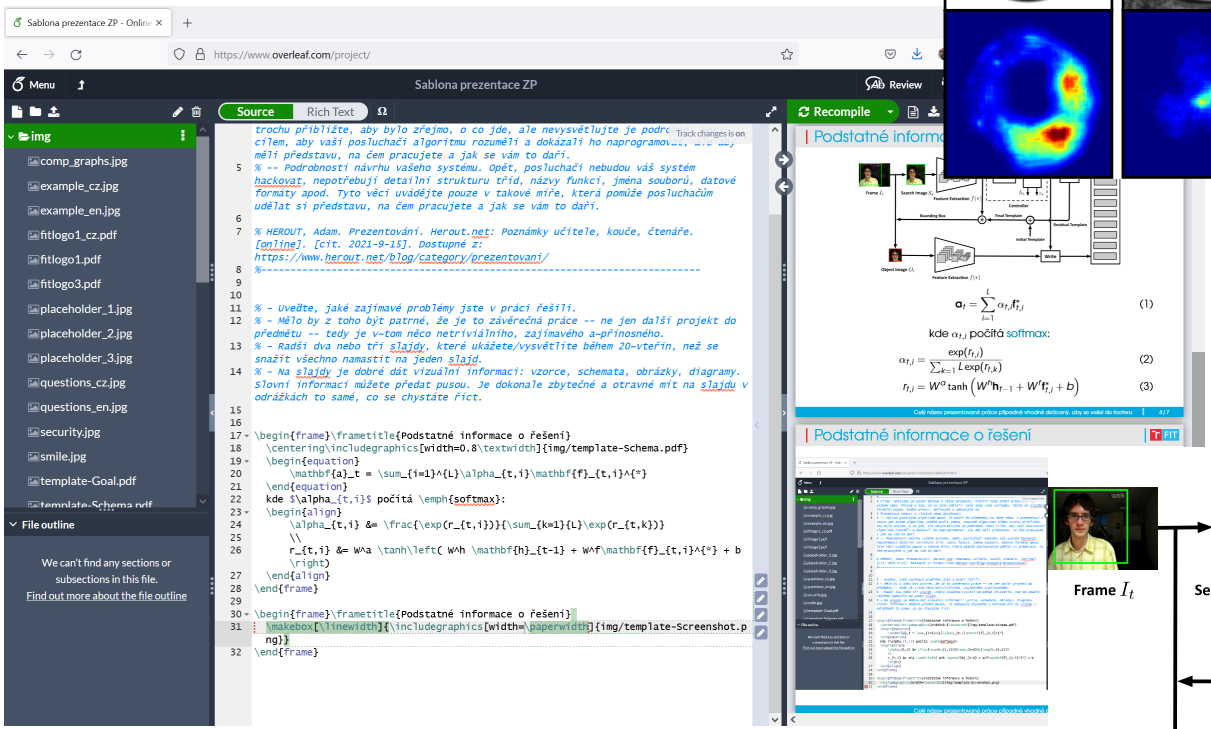
	AN	RN	EAO ↑	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	Speed
SiamFC	✔		0.188	-	-	86
SiamFC		✔	0.251	-	-	40
SiamRPN	✔		0.243	-	-	**200**
SiamRPN		✔	0.359	-	-	76
SiamMask-2B w/o R		✔	0.326	62.3	55.6	43
SiamMask w/o R		✔	0.375	68.6	57.8	58
SiamMask-2B-score		✔	0.265	-	-	40
SiamMask-box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

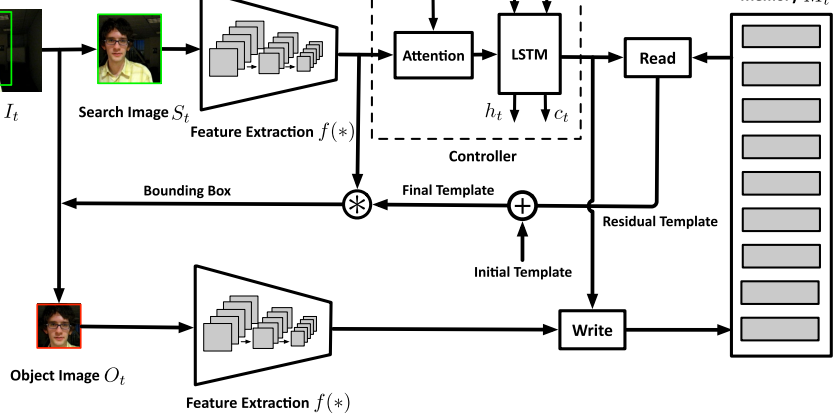




Děkuji za pozornost!







Otázky oponenta

- 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.