# Asociovanie videného obrazu a motorickej akcie na jeden pokus

**Andrej Lúčny**

Katedra aplikovanej informatiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského
KAI FMFI UK, Mlynská dolina, 842 48 Bratislava
lucny@fmph.uniba.sk

**Abstrakt**

Pojednávame o možnostiach učenia na jeden pokus, ktoré skúšame na reakcii robota na videný obraz. Vysvetľujeme, prečo je priame asociovanie obrazu s akciou neúčinné a prečo pomocou kóderov a dekodérov získaných hlbokým učením je možné túto úlohu úspešne vyriešiť. Pritom pre použitie asociácii využívame mechanizmus kľúčov a hodnôt, ktorý je jadrom tzv. transformátorov. Priestor príznakových vektorov – či už kódujúcich obraz alebo motorickú akciu – je totiž na rozdiel od priestoru vstupných obrazov a výstupných motorických akcií plynulý a každý jeho prvok zodpovedá ako tak rozumnej inštancii situácie, v rámci ktorej robot koná. Prínosom nášho príspevku je vhodný spôsob reprezentácie percepcie a akcie ako aj ich asociovania.

## 1 Úvod

V prírode pozorujeme rôzne podoby učenia sa. V Dennet 2008 sa rozlišuje medzi tzv. skinnerovským a popperovským typom mysle. Kým v prvom prípade prebieha zdokonaľovanie schopností postupným učením, na ktoré je potrebné podstúpiť veľa opakovaných pokusov, v druhom je možné aby proces učenia prebehol náhle, na základe jedinej skúsenosti s predmetnou situáciou. Hoci súčasné výdobytky umelej inteligencie nás vedia ohúriť svojimi schopnosťami, získavajú ich skinnerovským, nie popperovským spôsobom: v procese trénovania, pri ktorom je im každý zo vzorov správania predložený mnoho krát. Bolo by ale možné z nejakou takto vopred pripravenou sadou schopností rozbehnúť zdokonaľovanie popperovským spôsobom?

Táto otázka je naliehavá hlavne v mobilnej robotike, kde na robotovi máme už dnes k dispozícii technické možnosti spúšťania modelov hlbokého učenia, ale vykonať trénovanie alebo doladenie týchto modelov je kapacitne problematické.

V tomto príspevku sa zameriavame na zjednodušenú situáciu, kedy má takýto robot získavať schopnosť zvoliť správnu akciu na základe určitej percepcie. Konkrétne asociujeme motorickú akciu vedúcu k zaujatiu určitej pózy tela humanoidného robota (používame iCubSim z Vernon 2007) na základe videného obrazu pri tzv. imitačnej hre (Boucenna 2014) (kapitola 2). Pritom robot je vždy určitej situácii fyzicky vystavený, takže zapamätať si asociáciu medzi videným obrazom a motorickou akciou nie je problém. Problém je vedieť tieto asociácie neskôr použiť. Háčik spočíva v tom, že robot sa už nikdy viac nedostane do presne rovnakej situácie, než v akej si asociáciu zapamätal. Musíme teda navrhnúť nejaký šikovný mechanizmus pomocou ktorého robot reaguje na obraz, ktorý ešte nikdy nevidel a na ktorý sa to, čo pozná, len podobá. Toto riešime pomocou tzv. attention mechanizmu z Vaswani 2017, ktorý je bežnou súčasťou transformátorových hlbokých neurónových sietí, pričom my ho používame pomerne netradičným spôsobom (kapitola 3).

Avšak ani takýto mechanizmus nám reálne neumožní implementovať imitačnú hru, ak by sme asociovali priamo videný obraz a motorickú akciu. Ich dátové priestory sú príliš veľké, deravé a málo plynulé. Pokiaľ napríklad videnú situáciu reprezentujeme ako bod v mnohorozmernom priestore, tak:

- počet dimenzií tohto priestoru je daný súčinom počtu pixelov a farebných kanálov, čo sú rádovo státisíce
- v priestore máme všetky možné obrazové vstupy, pričom väčšinu z nich robot nikdy nemôže reálne uvidieť
- pri drobnej zmene situácie môže v takom dátovom priestore dôjsť k dramatickej zmene polohy reprezentujúceho bodu

Čo sa týka motorickej akcie:

- počet dimenzií bude rovný počtu stupňov voľnosti robota, t. j. počtu kĺbov, ktorých uhlami robota riadime (uhlové rýchlosti zanedbávame), čo síce nebude až tak veľa (ide o jednotky až desiatky), avšak:
- v priestore máme všetky možné variácie kĺbových uhlov, z ktorých väčšinu tvoria pózy, ktoré nie je vhodné zaujať (na rozdiel od človeka robot necíti, ktorá póza je mu príjemná a ktorá nekomfortná)
- pri drobnej zmene nastavenia stupňov voľnosti, väčšinou prichádza k drobnej zmene pózy, avšak vplyv zmeny rôznych stupňov voľnosti je veľmi rozdielny: hierarchicky vyššie stupne majú rádovo väčšie dopady na výslednú pózu

Prekonať zlé vlastnosti priestorov obrazov a póz nám však umožňuje hlboké učenie. Práve túto schopnosť jeho modelov považujeme za kľúčovú (kapitola 4).

Keď obraz a pózu spracujeme modelmi hlbokého učenia na zodpovedajúce príznaky, tieto dva príznakové vektory budeme vedieť nielen asociovať, ale túto asociáciu aj účinne použiť. Ich dátové priestory sú totiž:

- oveľa menšie (v prípade obrazu stovky – maximálne tisíce – dimenzií, v prípade pózy ide o jednotky)
- bez dier: každý príznakový vektor zodpovedá nejakému možnému videnému obrazu, respektíve rozumnej póze, ponajviac nejakej prechodnej forme medzi dvoma takými obrazmi či pózami
- plynulé: príznakové vektory zodpovedajúce postupnej zmene videnej situácie či zaujatej pózy predstavujú v priestore príznakov trajektóriu podobnú prechodu medzi počiatočným a koncovým stavom.

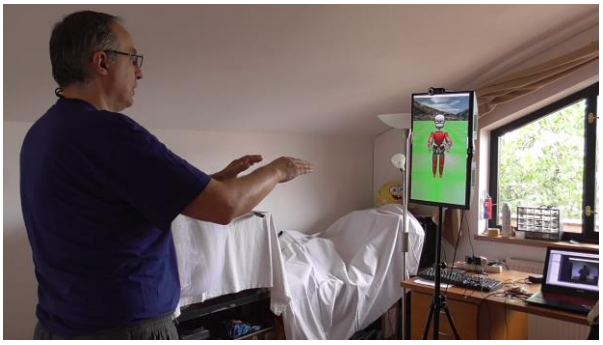
Vďaka týmto vlastnostiam príznakov sme schopní imitačnú hru implementovať (kapitola 5), čím implementujeme ukážku učenia sa robota na jeden pokus.

## 2 Imitačná hra

Cieľom imitačnej hry, na ktorej testujeme náš prístup, je naučiť robota imitovať človeka na základe toho, že človek imituje robota (obr. 1). Hra prebieha v dvoch fázach.

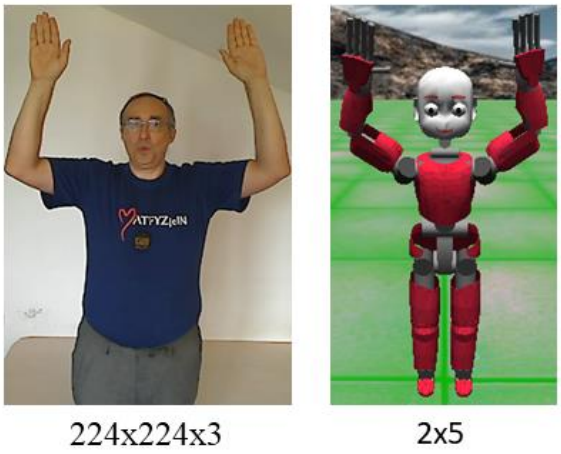
V prvej fáze robot vyzýva človeka, aby ho napodobňoval. Vytvára napríklad rôzne polohy rúk a človek ich napodobňuje svojím telom pred kamerou robota. To dáva šancu robotovi zapamätať si asociácie medzi pózami svojho tela a videným obrazom.

V druhej fáze robot napodobňuje človeka pomocou asociácií získaných v prvej fáze.



**Obr. 1:** Imitačná hra

Asociácie získané robotom predstavujú zoznam reprezentácii obrazu, ktorý robot vidí a pózy, ktorú robot zaujal. Obraz je vo svojej pôvodnej forme trojrozmernými poľom s rozmermi: výška, šírka a kanál, ktorého prvky sú intenzity od 0 do 255. V našej implementácii s farebným obrazom (3 kanály) používame rozlíšenie obrazu 224x224 pixelov. I pri tomto skromnom rozlíšení je počet dimenzií tohto dátového priestoru úctyhodných 224x224x3 = 150528. Pri póze uvažujeme len polohy rúk robota, čo je 2x5 = 10 stupňov voľnosti. V pôvodnej forme teda ide o vektor desiatich čísel zodpovedajúcim stupňom, ktoré normalizujeme do intervalu $<-1,1>$ (čo je vhodnejší rozsah pre spracovanie neurónovou sieťou). (Obr. 2) Spôsob reprezentácie oboch si môžeme zvoliť. Prínosom tohto príspevku je, že sme našli vhodný taký spôsob.




**Obr. 2:** Asociované dáta: obraz a póza

Technicky je náročné zabezpečiť, aby robot vo prvej fáze správne vystihol moment, kedy človek zaujme jeho pózu. To sme si zjednodušili tak, že to človek robotovi naznačí. Keďže jeho ruky sú zaneprázdnené zaujatím správnej pózy, využili sme na to detektor zapískania. Ten funguje na báze Fourierovej transformácie zvuku a analýzy amplitúdového spektra.

## 3 Mechanizmus asociovania

Nami používaný mechanizmus asociovania, známy pod – pre nás nie celkom vhodným – menom attention (pozornosť), pracuje s množinou $l$ párov kľúč - hodnota. Keď máme na vstupe dotaz $q$, namiešame dotaz z kľúčov $K$ a výstup vytvoríme ako analogickú zmes zo zodpovedajúcich hodnôt $V$, kde:

$$K = \begin{pmatrix} k_1 \\ k_2 \\ \dots \\ k_l \end{pmatrix} \qquad V = \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_l \end{pmatrix}$$

Všetky dotazy a kľúče sú vektory dimenzie $n$, takže $K$ je matica $l \times n$. Hodnoty a výstupy sú vektory dimenzie $m$, takže $V$ je matica $l \times m$. Najprv nájdeme také $c_i \in <0,1>$, že $\Sigma\, c_i\, k_i = pr_K(q)$, $\Sigma\, c_i = 1$ a $i=1,2,...l$, kde

$pr_K(q)$ je vektor podobný projekcii $q$ do podpriestoru generovaného kľúčmi $K$. Pritom chceme, aby $c_i$ vyjadrovalo podobnosť medzi kľúčom $k_i$ a dotazom $q$, takže ho môžeme odvodiť od skalárneho súčinu $q\ k_i$, úmernému uhlu, ktorý $q$ a $k_i$ zvierajú. Tieto podobnosti – pozitívne pre zhodné, nulové pre navzájom kolmé a záporné pre opačné vektory – však musíme dostať do $<0,1>$, čo nám dokáže zariadiť funkcia $softmax(x)_i = exp(x_i) / \Sigma_k\ exp(x_k)$. Koeficienty pomocou ktorých namiešame z kľúčov $k_i$ niečo podobné dotazu $q$, zvolíme preto ako:

$$c = softmax\left(\frac{qK^T}{d}\right)$$

kde $d$ je škálovací faktor, ktorým určujeme koľko namiešame z podobných kľúčov a koľko z odlišných. Čím menšia táto konštanta je, tým sa viac sa koeficienty blížia k tzv. one-hot kódu (jedna jednotka a ostatné nuly). Pre $d = 1/n$, kde $n$ je dimenzia kľúčov sa už prakticky vždy prikloníme k prevahe jedného kľúča, zatiaľ čo obľúbená hodnota $d = \sqrt{n}$ zabezpečuje, že vždy trochu miešame aj z ostatných kľúčov. To môže byť prínosom pre schopnosť asociačného mechanizmu vynájsť správnu odozvu aj pre také dotazy, ku ktorým podobný kľúč zapamätaný nemá, avšak dajú sa považovať za prechodnú formu medzi dvomi, či viacerými zapamätanými kľúčmi.

Keď už máme koeficienty zmesi $c$, ktorými sme približne vyjadrili dotaz podľa kľúčov, môžeme analogickým spôsobom zmiešať hodnoty $V$ na výstup $o = cV$. Takže úplná odpoveď asociačného mechanizmu **A** na dotaz $q$ je:

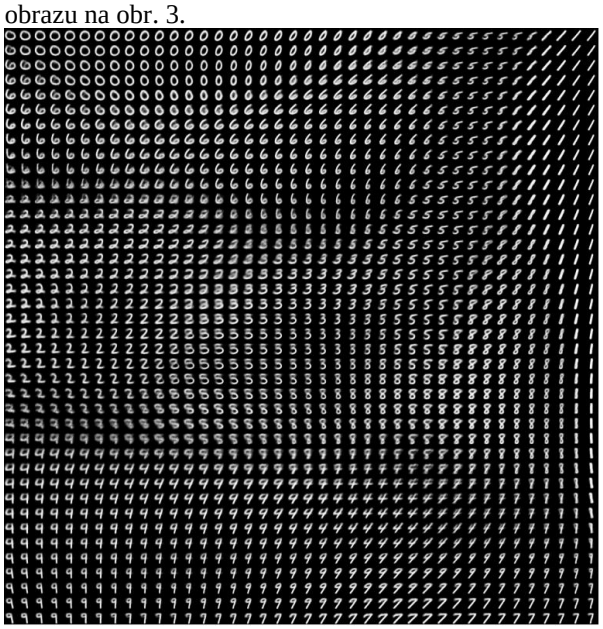
$$A(q, K, V) = softmax\left(\frac{qK^T}{d}\right)V$$

Sám o sebe však tento mechanizmus našu úlohu nevyrieši. Keby sme asociovali priamo obraz s akciou (čo je pri dimenziách $n=150528$ a $m=10$ technicky možné), dostávali by sme v konečnom dôsledku skoro vždy rovnakú odozvu, v dôsledku čoho by sa imitujúci robot sotva pohol.

## 4 Príprava modelov hlbokého učenia

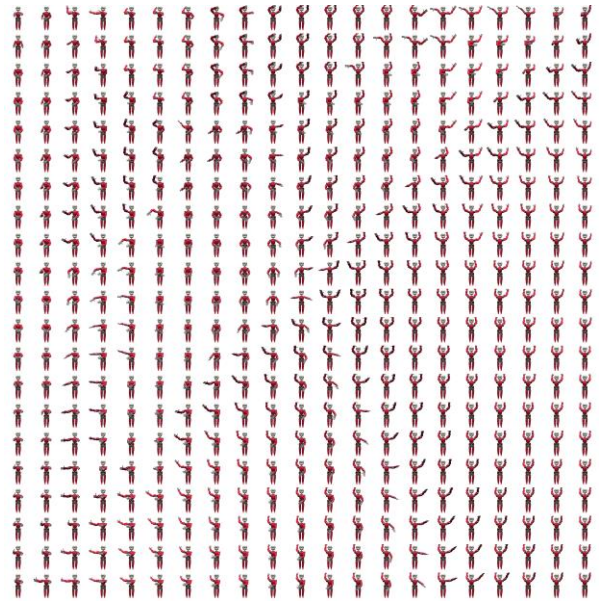
Na to, aby asociačný mechanizmus fungoval, potrebujeme, aby priestory kľúčov a hodnôt boli plynulé a neobsahovali žiadne diery. To sa dnes dá našťastie ľahko zariadiť, lebo práve tieto vlastnosti sú podstatou fungovania modelov tzv. hlbokého učenia. Náš robot do hry vstupuje s dvoma hotovými modelmi: kóderom obrazu a dekóderom pózy. Oba sa dajú získať bez potreby anotovania dát.

Ako kóder obrazu sme využili predtrénovaný vizuálny transformátor trénovaný samoučením (metódou DINO podľa Caron 2021, ktorej podstatou je predkladanie podobných a rozdielnych obrazov dvom kópiám tej istej siete, pričom v prvom prípade požadujeme rovnakú odozvu a v druhom rozdielnu), ktorý obraz kóduje do 384 príznakov. Urobiť si určitú predstavu o tom čo robí, je možné z oveľa jednoduchšieho príkladu analogického spracovania obrazu na obr. 3.



**Obr. 3:** Mapovanie ručne písaných číslic s rozlíšením 28x28 do latentného priestoru dimenzie 2. Podobné mapovanie robí DINO pre obraz pred kamerou, avšak v tomto prípade má latentný priestor 384 dimenzií a je ťažko si ho predstaviť

Dekodér pózy sme získali natrénovaním variačného autokódera (Kingma 2019) (vybrali sme lepší výsledok z viacerých trénovacích pokusov) z dátových vzoriek pohybov rúk robota do bodov v okolí robota. Tieto pohyby sme získali spätnou kinematikou (obr 4).



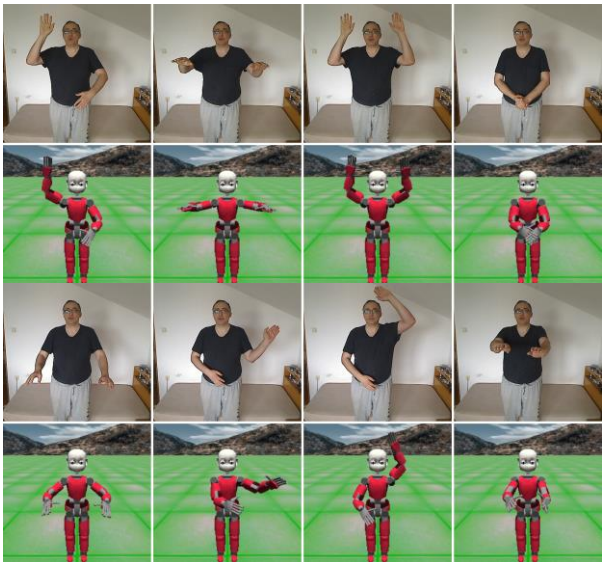
**Obr. 4:** Mapovanie póz robota 2x5 stupňov voľnosti do latentného priestoru dimenzie 2 pomocou konvolučného variačného autokódera.

Učenie robota imitácii bude spočívať v tom, že pomocou asociačného mechanizmu namapujeme jeden latentný priestor na druhý.

## 5 Implementácia imitačnej hry

V prvej fáze imitačnej hry robot z niekoľko vybraných príznakových vektorov pózy dekóduje pózu, zaujme ju, počká na signál od človeka, že aj on zaujal správnu pózu, zakóduje videný obraz do príznakov a uloží si oba príznakové vektory do zoznamu asociácii (príznaky obrazu budú kľúčom a príznaky pózy hodnotou). Na naučenie sa určitej sady polôh, stačí toľko asociácii, koľko je polôh, prípadne o niečo málo menej, keďže niektoré polohy je možné zložiť z ostatných.

V druhej fáze robot zakóduje obraz do príznakov, vypočíta z asociácii zodpovedajúci príznakový vektor pózy, dekóduje a zaujme zodpovedajúcu pózu (obr. 5). Pritom je schopný sa nielen prikloniť k niektorej zapamätanej póze, ale tieto aj vhodne skombinovať (táto schopnosť však silne závisí škálovacieho faktoru asociačného mechanizmu)[1].



**Obr. 5:** Priebeh imitačnej hry (druhá fáza)

Zaujímavou vlastnosťou tohto riešenia bolo, že ak človek robota v prvej fáze oklamal a miesto správnej pózy urobil niečo iné – napríklad mu ukázal nejaký objekt – robot sa túto neadekvátnu reakciu naučil[2].

## 6 Záver

Učenie asociovaním predpripravených modelov je prístupom, ktorý môže byť zaujímavý, ak potrebujeme, aby učenie prebehlo náhle a rýchlo. Užitočné môže byť hlavne pre mobilné roboty, ktoré na palubnom počítači vedia spúšťať modely hlbokého učenia, avšak nedisponujú kapacitou na ich trénovanie či doladenie. Vzbudzuje viacero otázok, ako dosiahnuť pri tomto spôsobe učenia čo najlepšiu kvalitu. Rezervy máme jednak v povahe kóderov, ktoré síce mapujú plynule, ale zaostávajú v uniformite, jednak v samotnom asociačnom mechanizme, ktorého rôzne obmeny sa chystáme skúmať.

Kódy a modely zdieľame na Github-e:
https://github.com/andylucny/cvae.git
https://github.com/andylucny/learningImitation.git

[1] viď video https://youtu.be/-3BVbU9BeRE
[2] viď video https://youtu.be/_CBnCOnWRdY


### Poďakovanie

Tento príspevok vznikol s podporou grantovej agentúry VEGA v rámci projektu 1/0373/23 a EU projektu TERAIS, č. 101079338.


## Literatúra


Boucenna, S., Anzalone, S., Tilmont, E., Cohen, D., Chetouani, M.: Learning of social signatures through imitation game between a robot and a human partner. *IEEE Transactions on Autonomous Mental Development 6(3)*, 213–225 (2014). https://doi.org/10.1109/TAMD.2014.2319861

Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the International Conference on Computer Vision. ICCV* (2021)

Dennett, D.C.: *Kinds of minds: towards an understanding of consciousness.* (1996) Weidenfeld & Nicolson, London

Kingma, D.P., Welling, M.: An introduction to variational autoencoders. (2019) *Foundations and Trends in Machine Learning 12(4)*, 307–392

Lucny, A.: Towards one-shot learning via attention. (2022) In: *CEUR Workshop Proceedings, ITAT 2022*. pp. 4–11. 3226

Šejnová, G., Štepánová, K.: Feedback-driven incremental imitation learning using sequential VAE. In: *2022 IEEE International Conference on Development and Learning (ICDL)*. pp. 238–243. IEEE (2022)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Lukasz Kaiser, Polosukhin, I.: Attention is all you need. (2017) In: *31st International Conference on Neural Information Processing Systems. ACM, Long Beach*

Vernon, D., Metta, G., Sandini, G.: The icub cognitive architecture: Interactive development in a humanoid robot. In: *2007 IEEE 6th International Conference on Development and Learning.* pp. 122–127 (2007). https://doi.org/10.1109/DEVLRN.2007.4354038