

Neuronové sítě – vektorové reprezentace

Ondřej Týbl

Statistický mikuláš

4. 12. 2025

Motivace

Pro účely zpracování, predikce a generování potřebujeme způsob, jak zakódovat vstup pomocí vektorů.

Motivace

Pro obrázky máme přirozeného kandidáta na takové kódování (vnoření), a sice *RGB reprezentace*.

Pro text přirozený kandidát **chybí**.

Motivace

Zabývejme se případem, kdy reprezentujeme slova pomocí vektorů v_j – text je pak sekvencí $v_1, v_2, \dots$.

Můžeme volit jednoduše

$$e_i = (0, \dots, 0, 1, 0, \dots, 0)^T \in \mathbb{R}^N,$$

kde N je **předem známá** velikost slovníku.

Motivace

Pak máme mnoho problémů:

- ▶ všechna slova jsou stejně vzdálená a svírají stejný úhel, nemáme zde žádný pojem *sémantické podobnosti*, neboť

$$\cos(\mathbf{e}_i, \mathbf{e}_j) = 0 \quad \text{a} \quad \|\mathbf{e}_i - \mathbf{e}_j\|_2 = \sqrt{2}$$

pro různá slova $i \neq j$,

- ▶ dimenzionalita reprezentace je dána přímo velikostí slovníku,
- ▶ $\{\mathbf{e}_j, \quad j = 1, \dots, N\}$ není uzavřená vzhledem k operaci sčítání ani násobení skalárem

Sémantická reprezentace

You shall know a word by a company it keeps.¹

¹Firth, J.R.: *Studies in Linguistic Analysis* (1957)

Sémantická reprezentace

Slova *učitel*, *žák*, *třída* vystupují v podobném kontextu – měly by mít podobnou reprezentaci.

Slovo *koruna* dostane význam teprve ve chvíli, kdy zjistíme kontext – měna nebo koruna stromu.

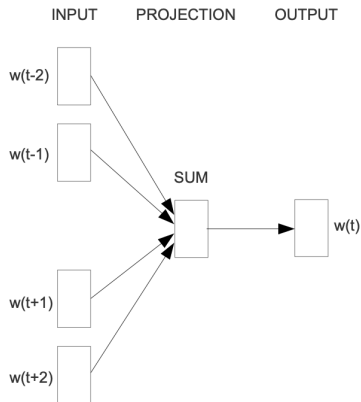
Sémantická reprezentace

Uvažujme hypotetickou reprezentaci j tého slova ve slovníku pomocí vektoru $w_j \in \mathbb{R}^d$ a označme $W = (w_1, \dots, w_N) \in \mathbb{R}^{d \times N}$. Poté přirozeně reprezentaci j tého slova obdržíme jako násobení We_j .

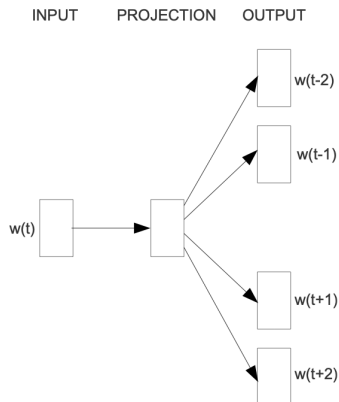
Matici W odhadneme na základě dat $v_1, v_2, \dots, v_K$ reprezentovaných kanonickými vektory pomocí optimalizace pro *predikci slova na základě svých sousedů*:

$$Wv_j \sim \frac{1}{4} (Wv_{j-2} + Wv_{j-1} + Wv_{j+1} + Wv_{j+2})$$

Sémantická reprezentace



CBOW



Skip-gram

Obrázek 1: Modely *Continuous Bag of Words* a *Skip-gram*, Mikolov et al.: *Efficient estimation of word representations in vector space* (2013).

Sémantická reprezentace

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Obrázek 2: Příklady sémantických vztahů; např. *Rome* je nejbližší vektor vektoru *Paris - France + Italy*, Mikolov et al.: *Efficient estimation of word representations in vector space* (2013).

Sémantická reprezentace

Několik poznámek

- ▶ v základní verzi byl použit text o 800 milionech slov (30 tisíc unikátních) při dimenzionalitě reprezentace 300,
- ▶ zpracování takto dlouhého textu je možné díky tomu, že není potřeba žádného lidského zpracování,
- ▶ dnes je standardní nepoužívat slova jako základní elementy, nýbrž *subwords* – nejčastěji používané sekvence v daném textu, výběr závisí na zvoleném algoritmu a jazyku
- ▶ (důležitý detail): technika negativního vzorkování je použita pro zrychlení výpočtů – pro každé slovo, zvýšíme pravděpodobnost skutečné třídy a snížíme pravděpodobnost pouze několika vybraným dalším třídám, třídy vybíráme na základě frekvenční analýzy

Sémantická reprezentace – dynamická verze

BERT (Bidirectional Encoder Representations from Transformers) poskytuje **kontextuální embedding**:

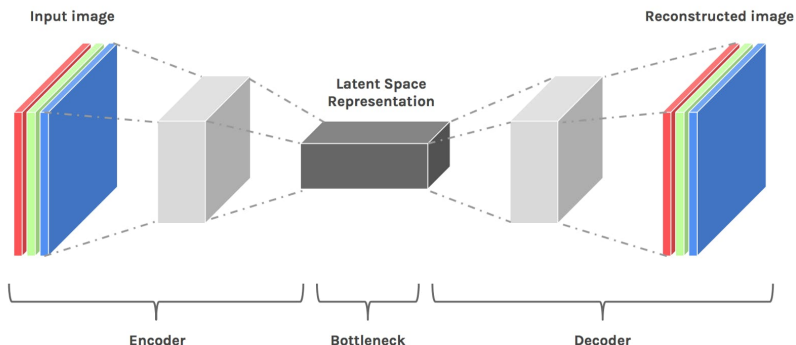
- ▶ jedno slovo nemá pevný vektor – jeho reprezentace se mění podle celého okolního textu,
- ▶ model čte větu obousměrně (doleva i doprava),
- ▶ embedding tedy zachycuje význam slova *v daném kontextu*,
- ▶ vhodné pro úlohy jako klasifikace vět, pojmenované entity, otázka–odpověď atd.

Příklad: „*koruna*“ má jinou reprezentaci ve větách „*vložit korunu do prasátka*“ a „*koruna stromu se třásla větrem*“.

Ukázka

Ukázka (klik)

Ve skutečnosti pro některé úlohy RGB reprezentace není vhodné kódování pro obrázky. Možným řešením je použití vnitřní reprezentace z tzv. *Variational Autoencoder*



Dovětek

Děkuji za pozornost