Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Michal Filippi
Název práce	Predikce sekundární struktury proteinu pomocí hlubokých neuronových sítí
Rok odevzdání	2017
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Mgr. Filip Matzner **Role** oponent
Pracoviště	KSVI

Text posudku:

Předkládaná práce se zaobírá predikcí struktury, kterou proteiny zaujímají v prostoru. Vzhledem k tomu, že experimentální analýza proteinů je velmi nákladná, je důležité umět strukturu proteinů alespoň co nejpřesněji predikovat. V této práci se k tomuto účelu používají hluboké neuronové sítě.

Práce je psaná srozumitelnou češtinou a v klíčových místech je text vhodně doplněn ilustracemi. Autor v průběhu celé práce cituje a používá state-of-the-art metody z posledních let.

Přesto, že práce nepřinesla výrazné zlepšení v samotné predikci sekundární struktury, nabízí mnoho zajímavých dílčích výsledků a rozsáhlou rešerši. Jedním ze zajímavých výsledků je, že konvoluční vrsva v state-of-the-art modelu DCRNN nepřináší téměř žádný užitek, stejně jako přidávání odhadu FCC mřížkové struktury na vstup sítě. Obě tato zjištění jsou nová a překvapivá. Za další zajímavou věc považuji, že naivní rekurentní síť neměla o tolik horší výsledky než moderní LSTM či GRU.

Autor zveřejnil veškerý software vytvořený v průběhu práce na svém GitHub účtu a umožnil tak snadné navázání na dosažené výsledky. Jeden ze zveřejněných programů je například nástroj FastProteinPSSM, který umožňuje řádově rychlejší výpočet PSSM oproti dosavadním metodám za cenu nepatrného zhoršení kvality. Autor dokonce ukazuje, že při použití v kombinaci s neuronovými sítěmi může toto zhoršení kvality donutit síť lépe generalizovat.

Kdybych práci musel něco vytknout, bylo by to následující. V textu se místy vyskytují drobné lingvistické chyby jako je špatná shoda podmětu s přísudkem či špatné skloňování (např. na půlstraně 28 je jich cca 6). Další výtky se týkají citací. U některých citací zcela chybí sborník, časopis nebo url (např. A. Cravens a C. Probert [2016] nebo Z. Li a Y. Yu [2016]), některé položky jsou nesourodé (např. někdy je u stránek napsáno slovo pages, někdy není), url je často jen odkaz do citačního manažeru nebo databáze doi a podobně. Citace jsou pravděpodobně vygenerovány generátorem bez dodatečné kontroly. Naštěstí nejsou chyby natolik závažné, aby citovaná práce

nešla najít. Další výtky se týkají obsahu. Na straně 31 je popis konvolučního bloku se vstupem velikosti [n,71] a třemi paralelními konvoluční vrstvami s 64 filtry, jejichž výstupy jsou spojeny za sebe. Výstup tohoto bloku má být velikosti [n,192], což není příliš v souladu s tím, co konvoluce dělá. Domnívám se tedy, že autor chtěl říci, že výstup konvolučního bloku je [n,71,192], přičemž dimenze o velikosti 71 je vnímána jako časová dimenze pro navazující RNN. Dále se domnívám, že zorné pole n sériově zapojených konvolučních vrstev se zorným polem r není r^n, jak je uvedeno na straně 33, ale spíše rn (pokud tedy není posun konvolučního okna stejně velký jako zorné pole). Poslední a možná trochu subjektivní poznámka je ta, že bych se nebál při definici známého výrazu uvést jeho originální anglickou verzi alespoň do závorky (např. dropout, bidirectional recurrent network, receptive field, ...).

Na závěr nutno dodat, že zmíněné výtky neubírají práci na zajímavosti, čtivosti nebo srozumitelnosti. Autor si chytře poradil s mnoha výzvami a provedl množství velmi dlouhých a náročných experimentů, jejichž výsledky shrnul v přehledných tabulkách. Práce svým rozsahem i obsahem splňuje podmínky závěrečné práce.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 27. 8. 2017

Podpis: