Strojový preklad pomocou hlbokých neurónových sietí

Oliver Tomko, Richard Záhumenský
Technická univerzita v Košiciach
TUKE
Košice, Slovenská republika
tomko.oliver@gmail.com
richard.zahumensky@student.tuke.sk

***Abstrakt*— Táto práca popisuje využitie hlbokých neurónových sietí v aplikáciách spracovania jazyka, najmä strojového prekladu. Venuje sa využitiu viacerých typov neurónových sietí ako sú RNN, BRNN, LSTM. Taktiež popisuje trendy a prax v spomínanej oblasti strojového prekladu.**

Kľúčové slová—deep learning, hlboké neurónové siete, RNN, BRNN, LSTM, rekurentné neurónové siete, Sequence-to-Sequence, Google translate, strojový preklad, spracovanie jazyka

I. Úvod

Pred niekoľkými rokmi bolo veľmi zložité preložiť text z neznámeho jazyka pomocou počítača. Použitie jednoduchých slovníkov s prekladom jednotlivým slov na slová cieľového jazyka bolo ťažké najmä z dvoch dôvodov: 1) čitateľ musel poznať pravidlá gramatiky a 2) bolo potrebné mať na pamäti všetky jazykové verzie pri prekladaní celej vety. Dnes sa však táto oblasť posunula míľovými krokmi dopredu vďaka využitiu umelej inteligencie, najmä neurónových sietí.[1]

II. História

Priblížme si kúsok históriu strojového prekladu. Prvé pokusy o strojový preklad začali krátko po 2. svetovej vojne. Očakávalo sa, že pre počítače táto úloha bude jednoduchá, avšak opak bol pravdou.

Prvý systém pre strojový preklad bol verejnosti predstavený 7.januára 1954 firmou IBM. Išlo o veľmi jednoduchý systém, ktorého slovník obsahoval iba 250 slov a pri prekladaní prekladal len 49 zvlášť vybraných viet z ruštiny do angličtiny.

Úspech tohto systému však motivoval pokračovať v tejto oblasti ďalej a najmä za účelom získavania výhod v studenej vojne. Boli vyvíjané systémy, ktoré prekladali ruské vedecké texty do angličtina, aby sa spoľahlivo monitoroval pokrok druhej strany vo vede. Kvalita týchto prekladov však bola veľmi nízka a musela sa do prekladať profesionálom.

V ďalších desaťročiach bola do oblasti strojového prekladu investovaná veľká dávka úsilia ako z komerčnej sféry, tak i akademickej.

Využívali sa metódy ako pravidlový strojový preklad, štatistický strojový preklad a hybridný strojový preklad.

Dnes môžeme dosiahnuť pokroky najmä pomocou využitia metód umelej inteligencie.[4]

Deep learningové aplikácie v rozoznávaní reči sa objavili v 90 rokoch minulého storočia. Prvý vedecký článok o využívaní neurónových sietí v strojovom preklade sa objavil roku 2014 a odštartoval následný progres v tejto oblasti.[5]

III. Problémy Pri preklade

Ak sa nástroj strojového prekladu (napr. Google Translate) pokúšal zachovať preklady aj pre krátke vety, nefungovalo to z dôvodu veľkého množstva možných variácií. Najlepšou myšlienkou je naučiť algoritmus gramatickým pravidlám a prekladať ich podľa nich.

Ak ste niekedy skúšali cudzí jazyk, viete, že vždy existuje veľa výnimiek z pravidiel. Keď sa snažíme zachytiť všetky tieto pravidlá, výnimky a výnimky z výnimiek v programe, kvalita prekladov sa rozpadne.

Moderné systémy strojového prekladu preto používajú odlišný prístup: pravidlá dolujú z textov analýzou obrovského možstva súborov a dokumentov.[1]

Ďalším problémom s ktorým sa pri preklade môžeme stretnúť je problém poradia sekvencií. Pri väčšine prípadov kde pracujeme so sekvenciami informácií záleží na ich poradí. Ak je poradie jednotlivých častí zmenené, celková informácia má buď pozmenený zmysel alebo nám vôbec žiadny význam nedáva.

S týmto problémom sa môžeme stretnúť pri spracovaní jazyka, ale aj pri spracovaná sekvencií udalostí, spracovaní sekvencií genómov atď.[3]

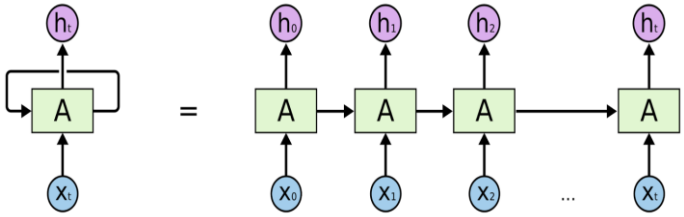
IV. Rekurentné neurónové siete

V našom prípade strojového prekladu, bude Rekurentná sieť prvým kúskom v skladačke úspešného riešenia. Pomocou nej máme prístup k dátam v predošlých časových okamihoch.

Pri rekurentnej neurónovej sieti je každá skrytá vrstva charakterizovaná vlastnými váhami a biasmi – ktoré môžeme použiť ako vstup.

Tieto nezávislé vrstvy potom zaobalíme/skombinujeme do rekurentnej vrstvy.

Rekurentný neurón drží stav o predošlého vstupu a kombinuje ho s aktuálnym vstupom. Teda dostávame vzťah medzi aktuálnymi a minulými stavmi, ktorý nám pôsobí na výsledok.



Rozvinutá rekurentná neurónová sieť

V. UČENIE RNN

Pri RNN nemusíme mať vstupy pri každom časovom kroku. Pri Forward Propagation data vstupujú do siete a idú dopredným smerom s každým časovým krokom.

Pri Backpropagation ideme akokeby „späť v čase“ aby sme zmenili váhy, takže sa to nazýva „Back propagation trought time“ – BPTT.

Chyba je vypočítaná ako strata entropie medzi dvoma hodnotami a to aktuálnym ty a predikovaným tȳ.

$Et(\bar{y}t,yt) = - \bar{y}t \log(yt)$

$E(\bar{y},y) = - \sum \bar{y}t \log(yt)$

Kroky pri spätnej propagácii chyby môžeme popísať nasledovne:

1. Krížová entropia chyby je vypočítaná použitím predikovaného výstupu a aktuálneho výstupu
2. Gradient je vypočítavaný pre každý časový krok vzhľadom na váhy
3. Gradient je kombinovaný pre všetky časové okamihy
4. Váhy sú aktualizované pre rekurentné neuróny aj pre fully connected vrstvy

Algoritmus backpropagácie RNN sa podobá na algoritmus klasickej neurónovej siete s výnimkou toho, že kombinujeme gradienty chyby pre všetky časové okamihy.[3]

VI. LSTM NEURONOVÉ SIETE

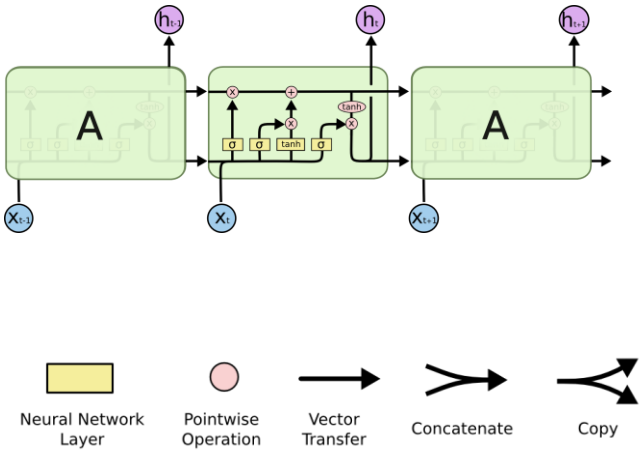
Klasické hlboké neurónové siete môžu dosiahnuť vynikajúce výsledky vo veľmi zložitých úlohách (rozpoznávanie reči / vizuálneho objektu), ale napriek ich flexibilite ich možno aplikovať len na úlohy, kde vstup a cieľ majú pevnú dimenziu.[1]

Tu prichádzajú na rad siete Long Short-Term Memory (LSTM), ktoré nám pomáhajú pracovať so sekvenciami, ktorých dĺžku nemôžeme a priori poznať.

Navyše sú aj vhodné pre modelovanie závislostí, ktoré majú veľký rozostup v čase. Je to spôsobené tým, že berú do úvahy viac predošlých stavov nielen jeden.

Napríklad pre generovanie Ht používame nielen Xt, ale všetky predchádzajúce vstupné hodnoty X

LSTM sú vlastne špeciálnym druhom rekurentnej neurónovej siete (RNN), schopnej učiť sa dlhodobým závislostiam. A kedže všetky RNN vyzerajú ako reťazec opakujúcich sa modulov, nie je to výnimkou ani pre LSTM. [1]



Hlavnou myšlienkou tejto siete je, že uchováva svoj stav. V LSTM máme brány – „gates“ pomocou ktorých môžeme buď vymazať alebo pridať informáciu do bunky.

Táto vrstva tvoriaca bránu môže nadobúdať hodnoty medzi 0-1. Hodnota 0 hovorí o tom, že nevpúšťame žiadne informácie ďalej, hodnota 1, že púštame všetko.

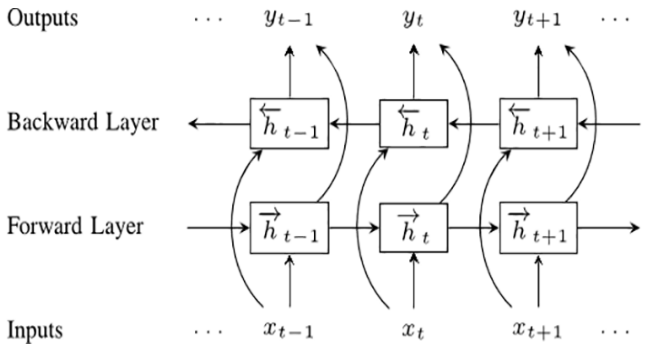
Prvým krokom LSTM je rozhodnúť, ktoré informácie vyhodíme zo stavu bunky a ktoré ponecháme. Napríklad pri spracovaní jazyka môžeme nahradiť rod minulého slova rodom nového slova. Možnosť ukladať viac informácii v bunkách nám lepšie pomáha modelovať jazyky.

Output z bunky je závislí na stave bunky avšak je to filtrovaná verzia. Sigmoidálna vrstva nám určí, ktorá časť stavu bunky ide na výstup. Stav bunky pustíme cez funkciu tanh a vynásobíme výstupom sigmoidálnej brány, aby sme na výstupe mali len časti ktoré potrebujeme.

Existuje viac variánt LSTM, ktoré môžeme ľubovoľne pozmeniť a prispôsobiť pre naše konkrétne úlohy.[6]

VII. Obojsmerné RNN

Ďalším krokom vpred pri strojovom preklade sú obojsmerné rekurentné neurónové siete (BRNN). Čo robí BRNN, je rozdelenie neurónov normálneho RNN do dvoch smerov. Jeden smer je pre pozitívny čas, alebo dopredné stavy. Druhý smer je pre záporný čas alebo pre spätné stavy. Výstup týchto dvoch stavov nie je pripojený na vstupy opačných smerov.



Obojsmerné rekurentné neurónové siete

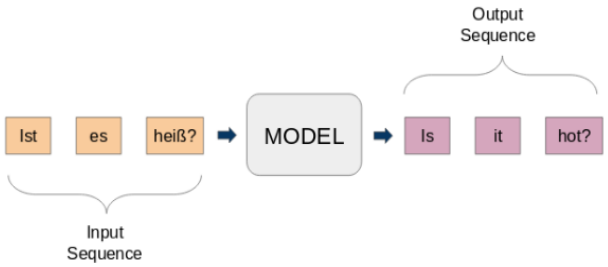
Aby sme pochopili, prečo BRNNs môžu fungovať lepšie ako jednoduché RNN, predstavte si, že máme vetu 9 slov a chceme predpovedať 5. slovo. Môžeme to spoznať buď iba prvé 4 slová, alebo prvé 4 slová a posledné 4 slová. Samozrejme, že kvalita v druhom prípade by bola lepšia.[1]

VIII. Sekvenčné modelovanie

Teraz sme pripravení prejsť na sekvenciu sekvenčných modelov (tiež nazývaných seq2seq). Základný model seq2seq sa skladá z dvoch RNN: sieť kódovačov, ktorá spracováva vstup a sieť dekodérov, ktorá generuje výstup.[1]

Sequence-to-Sequence (seq2seq) modely sa používajú pre rôzne úlohy NLP, ako je napríklad sumarizácia textu, rozpoznávanie reči, modelovanie sekvencií DNA, okrem iného. Naším cieľom je prekladať dané vety z jedného jazyka do druhého.

Tu sú oba vstupy aj výstupy vety. Inými slovami, tieto vety sú sledom slov idúcich dovnútra a von z modelu. Toto je základná myšlienka modelovania Sequence-to-Sequence. Obrázok nižšie sa snaží vysvetliť túto metódu.



Typický model seq2seq má 2 hlavné komponenty kodér a dekóder. Obe tieto časti sú v podstate dve rôzne modely rekurentnej neurónovej siete (RNN) kombinované do jednej obrovskej siete.[2]

Prípady použitia sekvenčného modelovania:

- Rozpoznávanie reči
- Name Entity / Subject Extraction na identifikáciu hlavného predmetu zo súboru textu
- Vzťah Klasifikácia medzi vzťahmi tagov medzi rôznymi entitami označenými vo vyššie uvedenom kroku
- Zručnosti chatbota mať schopnosť konverzácie a spojiť sa so zákazníkmi
- Sumarizácia textu na vytvorenie stručného zhrnutia veľkého množstva textu
- Systémy odpovedania na otázky
- Klasifikácia sentimentu - môžeme klasifikovať tweety, komentáre, správy na sociálnych sieťach do tried napr. pozitívne/negatívne

IX. Oblasti spracovania jazyka

Hlboké neurónové siete sú široko využiteľné vo všetkých oblastiach spracovania jazyka, nielen pri preklade.

Tieto oblasti môžeme rozdeliť na viac smerov. Uvedieme si teraz niektoré z nich.

NLP: Natural Language Processing: tento smer spracovania jazyka sa zaoberá analýzou textu. Snaží sa v ňom identifikovať lingvistické kategórie – ako napr. slovné druhy, slovnú zásobu atď a klasifikovať ich.

NLU: Natural Language Understanding: je väčšinou považovaný za podmnožinu NLP, ktorý sa zaoberá významom slov. Ponúka nám pochopenie ako sú slová využívané v kontexte situácií, ktoré napríklad zahŕňajú sarkazmus, ironiu, citlivost alebo humor.

Natural Language Generation: NLG sa nezaoberá analýzou sémantického výzamu, ale jeho vytváraním generáciou. Taktiež sa používa na zovšeobecňovanie informácií. Pomocou neho

vieme napríklad prezrieť 100 dokumentov a dostať sumárne informácie o týchto dokumentoch.

Natural Language Interaction: NLI je kombináciou predošlých oblastí a systémov, ktoré vytvárajú system, ktorý vie interagovať a komunikovať s používateľom. Je schopný generovať odpoveď.[7]

X. GOOGLE TRANSLATE

Architektúra NN bola postavená na modeli seq2seq, ktorý sme už študovali.

Jedinou výnimkou je, že medzi snímačom a dekodérom je 8 vrstiev LSTM-RNN, ktoré majú zvyškové spojenia medzi vrstvami s niektorými vylepšeniami pre presnosť a rýchlosť.

Systém vyžaduje „token“ na začiatku vstupnej vety, ktorá určuje jazyk, do ktorého sa pokúšame preložiť frázu.

To zlepšuje kvalitu prekladu a umožňuje preklady aj medzi dvomi jazykmi, ktoré systém ešte nevidel, metóda nazývaná „Zero-Shot Translation“. [1]

XI. OHODNOTENIE KVALITY PREKLADU

Jedna veta sa dá preložiť viacerými spôsobmi. Ako určíme, ktorý kandidát na preklad je vhodnejší?

Nie je to triviálny problém. Existuje množstvo prístupov, ktoré čiastočne riešia tento problém, ale najpopulárnejšou metrikou je BLEU – tkz. (bilingual evaluation understudy). Princíp spočíva v tom, že porovnáme kandidátov strojového prekladu s referencovanými prekladmi od profesionálnych prekladateľov. Tieto preklady existujú pre bežné vety využívané v jazykoch. Metóda BLEU nám vyhodnotí kvalitu pomocou určenia počtu slov, ktoré majú prienik v oboch vetách (nezávisle na pozícii).

Preto pri algoritmoch strojového prekladu ide vždy o dolovanie veľkého množstva znalostí z dát, najčastejšie z textových dokumentov.[1]

XII. NAJMODERNEJŠIE TRENDY

Od roky 2016 väčšina najlepších systémov pre strojový preklad využívalo neurónové siete. Prekladacie služby od Googlu, Microsoftu, Yandexu a PROMT-u dnes využívajú strojový preklad pomocou neurónových sietí.

Google uprednostňuje svoj GNMT – Google Neural Machine Translation pred technológiami, ktoré využíval predtým – štatistické metódy.

Skupina z Harvardu, ktorá sa týmto technológiám venuje uviedla svoj open souce systém OpenNMT.[5]

Myslíme si, že vývoj v tejto oblasti je veľmi perspektívny a bude pokračovať. Perspektívy sú pozorovateľné v kombinácii rôznych typov hlbokých neurónových sietí, ale aj vo vyvíjaní ich nových typov a modifikácií.

Taktiež si myslíme, že by vývoju mohli prospieť hybridné systémy, založené na kombinácii neurónových sietí a fuzzy systémov. Využívanie fuzzy množín má blízko k modelovaniu jazykových procesov, pretože každé slovo odzrkadľuje realitu s určitou hodnotou príslušnosti. Slová sú „fuzzy“, neodrážajú realitu presne a exaktne. ϖν⟨ ϖεχ,

Prínosom v tejto oblasti môže byť i štúdium lingvistiky a neurolingvistiky a vytváranie topológií hlbokých neurónových sietí inšpirovaných biologickými modelmi.

POUŽITÉ ZDROJE

[1] Daniil Korbut, “ Machine learning translation and the GOOGLE translate algorithm” .

[2] Prateek Joshi.,” NLP tutorial on neural machine translation – the technique powering Google transate ”.

[3] Disha Shree Gupta, “Fundamentals of Deep Learning – Introduction to Recurrent Neural Networks”

[4] Strojový preklad: wikipedia

[5] Neural machine translation: wikipedia

[6] Christopher Olah „Undestanding LSTM Neural Networks “

[7] Jelani Harper: Trends in Natural Language Processing