Valodas mašīnmācīšanās no nemarķēta teksta

Pēteris Paikens
peteris@ailab.lv

LU 77. konference

2019.02.21

Mākslīgā intelekta laboratorija
LU MII

Pārneses mācīšanās

Lietojam vienā uzdevumā gūto pieredzi citur

- Mācoties jaunu valodu, lietojam zināšanas par vecajām
- Mācoties braukt ar riteni, lietojam līdzsvara izjūtu

Apmācīt MM modeli vienam uzdevumam, bet lietot citam

- Pielāgošana nozarei (*Domain adaptation*)
- Pielāgošana uzdevumam (*Task adaptation*)
- Datu reprezentācijas iemācīšanās

Valodas apstrādes "ImageNet brīdis"

ImageNet – 2009 gada datu kopa attēlu klasificēšanai

- Daudz kvalitatīvu datu skaidri definētai problēmai
- Pāris gadu laikā daudzi aizvien labāki risinājumi
- Pamats dziļās mašīnmācīšanās revolūcijai
- Tiek lietots arī **visiem citiem** attēlu uzdevumiem

Kā iegūt šādu rezultātu valodas apstrādē ?

Kas tam bija vajadzīgs?

1. Daudz datu

   1.3 mlj attēlu

2. Arhitektūra, kas piemērota daudziem uzdevumiem

   Konvolūciju neironu tīkli

3. Uzdevums, kas piespiež iemācīties būtisko

   1000 sarežģītu kategoriju klasifikācija

4. Spēja efektīvi apmācīt sistēmas

Mēģinājums #1 – n-grammu modeļi

1. Daudz datu

   1995 British National Corpus, 100 mlj vārdu

2. Arhitektūra, kas piemērota daudziem uzdevumiem

   …. nē.

3. Uzdevums, kas piespiež iemācīties teksta jēgu

   Paredzēt vārdu virknes varbūtību

4. Spēja efektīvi apmācīt sistēmas

   Jā, pat ļoti lielas

#2 – Collobert/Weston 2008

1. Daudz datu

   1 mlj vārdu anotēts, 631 mlj vārdu vikipēdija

2. Arhitektūra, kas piemērota daudziem uzdevumiem

   Jā – Vārdu vektortelpa un konvolūciju tīkli

3. Uzdevums, kas piespiež iemācīties teksta jēgu

   Jā – POS, NER, *chunking,* SRL (PropBank) un valodas modelis

4. Spēja efektīvi apmācīt sistēmas

   Nē, ne 2008. gadā

#3 – vārdu jēdzientelpa, word2vec

1. Daudz datu

   6 mljrd vārdu Google News korpuss

2. Arhitektūra, kas piemērota daudziem uzdevumiem

   Daļēji

3. Uzdevums, kas piespiež iemācīties teksta jēgu

   Vai vārds "iederas" kontekstā

4. Spēja efektīvi apmācīt sistēmas

   Jā!

#4 – *Unsupervised sentiment neuron*

1. Daudz datu

   82 mlj dokumentu
2. Arhitektūra, kas piemērota daudziem uzdevumiem

   Nē – vienvirziena mLSTM
3. Uzdevums, kas piespiež iemācīties teksta jēgu

   Paredzēt nākamo burtu no iepriekšējiem …
4. Spēja efektīvi apmācīt sistēmas

   Jā!

Pēdējā gada pārmaiņas

1. ULMFiT (Universal Language Model Fine-tuning)
2. ELMo (Embeddings from Language Models)
3. GPT (Generative Pre-Training)/OpenAI Transformer
4. BERT (Bidirectional Encoder Representations from Transformers)
5. GPT-2 (Generative Pre-Training)

Kā tās darbojas?

Apmācam **dziļu** tīklu uz daudz nemarķēta teksta

- word2vec dod pirmo slāni, pārējos apmāca
- BERT, ELMo, utt dod visu sistēmu, apmāca tikai pēdējo slāni "atbildes nolasīšanai"

Triki, lai vajadzētu ņemt vērā attālu kontekstu

- pilns teksts, nevis teikumi vai pat rindkopas
- teikumu secības prognozēšana

Kā ir tagad?

1. Daudz datu

   BERT – 3.3 mljrd vārdu; GPT2 – 8 mlj dokumentu, 40 GB

2. Arhitektūra, kas piemērota daudziem uzdevumiem

   Jā – Transformer (*self-attention*) slāņi

3. Uzdevums, kas piespiež iemācīties teksta jēgu

   Jā – Valodas modelēšana + atsevišķi triki

4. Spēja efektīvi apmācīt sistēmas

   Jā, ar nosacījumu – ja ir lieli skaitļošanas resursi, tad var ātri

Ko šīs sistēmas spēj?

1. Pārspēj labākos rezultātus daudzos uzdevumos
   - Minimāla adaptācija bez iedziļināšanās problēmā
2. Bezparaugu mācīšanās (*Zero-shot learning*)
   - Gatavam modelim "uzdodam jautājumu" ar jaunu tēmu
3. Nav sasniegušas robežu
   - Vairāk datu – joprojām labāk
   - Vairāk parametru – joprojām labāk

Sekas valodas apstrādei vispār

1. Aizvietos *word embeddings* gandrīz visās sistēmās
   - Publiski pieejamas (angļu val.)
   - Ātrdarbības un veiktspējas ierobežojumi
2. Empīriski demonstrē to, ka valodā viss ir saistīts
   - Valodnieciskas parādības var iemācīties no teksta
3. Pieaug "ieejas biļetes" cena aparatūras ziņā
4. Pieaug nozīme rakstītās valodas krājumiem
   - Atpaliks valodas ar maz runātājiem, maz literatūras

Sekas latviešu valodas apstrādei

1. Pasaulē veido "daudzvalodu" sistēmas
   - Vispārīgi risinājumi bez adaptācijas katrai valodai
   - Risinājumu arhitektūra atbilst angļu un ķīniešu valodām
2. Iespējas adaptēt morfoloģiski bagātām valodām
   - Metodes ir, taču tās jāpielieto pašiem
3. Marķētu resursu izmērs kļūst mazāk svarīgs
   - Marķēt daudz resursus ir dārgi – bet vajag kvalitāti
4. Vajag **lielus**, labus korpusus

Sekas mākslīgā intelekta jomai

- Reinforcement learning (cherry)
- Supervised learning (Chocolate)
- Unsupervised/Predictive learning (Cake)
  - Generative adversarial nets (GAN)

Sekas mākslīgā intelekta jomai

- Paredzēt "kas būs?" liek iemācīties struktūru
  - Dialoga paredzēšana – atbilde uz jautājumu
- Attēlu paredzēšana – nākamais "kadrs" ?
  - Fizika, apkārtējās pasaules uzvedība
  - Piemēri par "spēļu pasauļu" prognozēšanu
  - Taustes, skaņas paredzēšana – reakcija uz rīcību
- Pastiprinātā mācīšanās – nevis galvenais datu avots, bet atgriezeniskā saite par augsta līmeņa mērķi

Paldies par uzmanību!
Jautājumi?

2008 R. Collobert, J. Weston "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning" https://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf

2013 T. Mikolov et al "Efficient Estimation of Word Representations in Vector Space" https://arxiv.org/abs/1301.3781

2017 A. Radford et al "Learning to Generate Reviews and Discovering Sentiment" https://arxiv.org/abs/1704.01444

2017 A. Vaswani et al "Attention Is All You Need" https://arxiv.org/abs/1706.03762 (*Transformer* modeļi)

2018 J. Howard, S. Ruder "Universal Language Model Fine-tuning for Text Classification" https://arxiv.org/abs/1801.06146

2018 M.E. Peters et al "Deep contextualized word representations" https://arxiv.org/abs/1802.05365 (ELMO sistēma)

2018 A. Radford et al "Improving Language Understanding by Generative Pre-Training" https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

2018 J. Devlin et al "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" https://arxiv.org/abs/1810.04805

2019 Yoav Goldberg "Assessing BERT's Syntactic Abilities" https://arxiv.org/abs/1901.05287

2019 A. Radford et al "Language Models are Unsupervised Multitask Learners" https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (GPT-2 sistēma)