A(sk) M(e) A(nything)

LING 575K Deep Learning for NLP

Shane Steinert-Threlkeld

May 17 2021

Announcements

- HW5 grades posted: great job!
- HW6 ref code available
- Packing and padding in HW7 (see these docs):
  - Saves computation by not running padding tokens through the RNN
  - Helpful visualization of packing: https://stackoverflow.com/a/56211056

```
# pack the sequences first; this prevents the RNN from computing on the padded tokens
packed_sequence = nn.utils.rnn.pack_padded_sequence(
    embeddings, lengths, batch_first=True, enforce_sorted=False
)
# get packed sequence out of LSTM
lstm_output, _ = self.lstm(packed_sequence)
# unpack sequence
# [batch_size, max_seq_len, hidden_dim]
lstm_output, _ = nn.utils.rnn.pad_packed_sequence(
    lstm_output, batch_first=True, padding_value=self.padding_index
)
```

Today's Plan

- Wrap-up interpretability/analysis
  - Attention-based methods
  - Adversarial data
- AMA / general discussion [thanks everyone!]
- Next three meetings: guest lectures
  - Ethics, fairness, limitations [Angelina McMillan-Major]
  - Multimodal NLP [Yonatan Bisk]
  - Low-resource / Multilingual NLP [C.M. Downey]

Q1: Reading Papers

- “How many papers a month do you (Shane) read a month to keep up to date with NLP deep learning? How many papers a month should a PhD. student read to do the same?”

Q1: Reading Papers

- TBH, I don't keep track, so I can't give an exact number.
  - But: do block out special time in calendar just for reading.
- But some thoughts:
  - It's *impossible* to read everything, so don't try :)
  - There are several "stages" of reading a paper.
  - Don't try to read line-by-line for complete understanding in the first pass.
    - Abstract + discussion: general idea of what happens
    - Methods/models: what did they really do
    - Deep dive: if I want to build on/extend the work

Q1: Reading Papers

- How to filter the noise
  - Google Scholar / Semantic Scholar searches
    - Also: chasing references from and citations to papers you know/like
    - Alerts for keywords, particular authors, …
  - ACL Anthology
  - arxiv-sanity.com
  - I like Sebastian Ruder's newsletter as well: https://newsletter.ruder.io/
  - Twitter*
    - *: I *think* it's a net-positive for research, but definitely not essential

Q2: Character-level vs. Subword

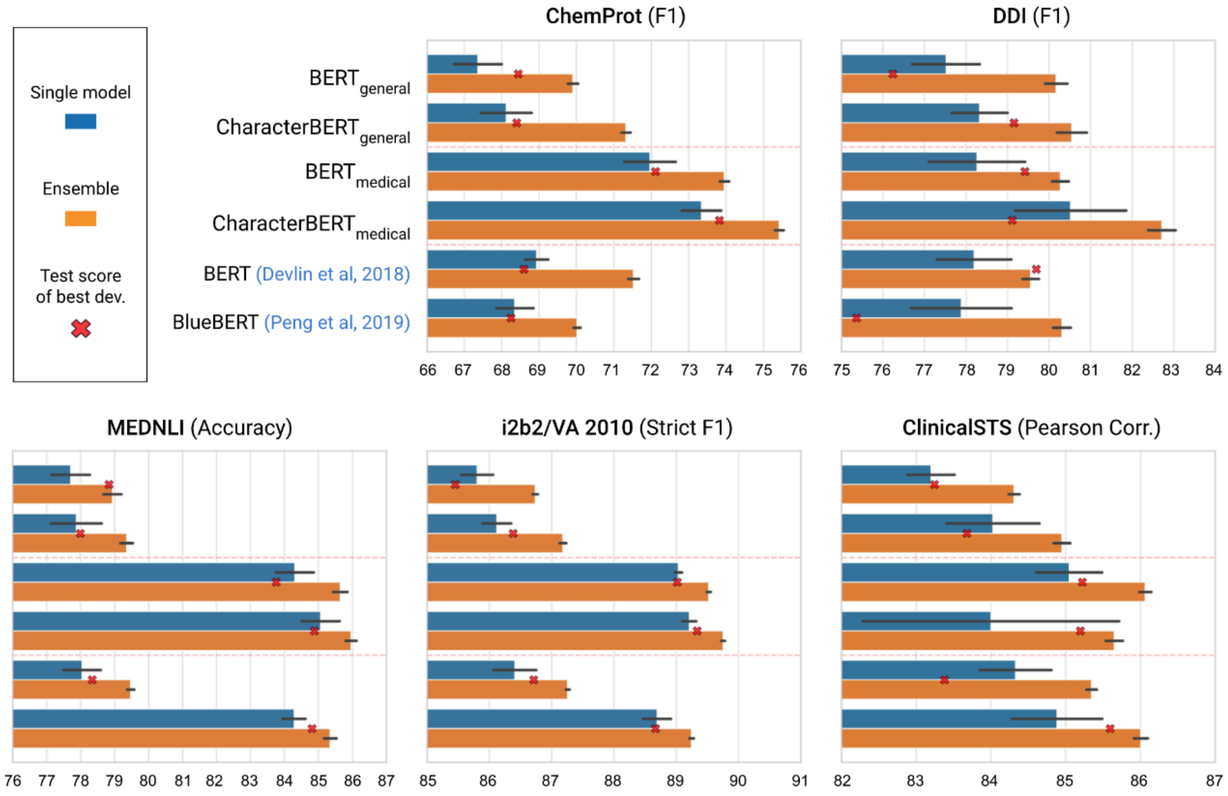
- “I remember reading a paper about testing character-level tokenization for BERT to improve performance on more domain-specific data (e.g. medical terminology, etc), and in the paper it said that ELMo was character-level based as well. I always thought of BERT as the "next step" from ELMo, but that seems like a pretty big distinction in basic model specifics that would influence behavior. Not sure what the question is here - I suppose I was wondering about if there were intermediate steps led people changing to WordPiece/BPE as the default BERT tokenization scheme.”

Q2: Character-level vs. Subword

- ELMo: still a *word-level* LM
  - But word representations are built by a *character-level* CNN
- So char-CNN and subword/BPE are two ways of handling/avoiding OOV
- Both approaches pre-date:
  - Char-CNN: e.g. Exploring the limits of language modeling
  - WordPiece: Google NMT

Q2: Character-level vs. Subword

- CharacterBERT (I think what the question was about :)): BERT, but char-CNN instead of WordPiece
- Performance seems to improve
  - Would like to see (i) basic MLM performance, (ii) non-medical transfer experiments
- Slightly slower at inference [CNN vs. lookup table]



Q3: LLMs and Paywalls

- “It seems wild to me that GPT-3 hasn't been released, and that you have to pay to be able to use it. At the same time, it seems like a logical outcome of the current trends in neural language modelling; only big corporations have the compute power to train these larger and larger LMs, and big corporations exist to make profit above all else. What do you think the future holds for large language models? Do you think future models will be behind a paywall, the way GPT-3 is? At what point are these big models just too big to be useful?”

Q3: LLMs and Paywalls

- NB: more next week :)
- Editorial remark: I find “Open”AI’s transition into a semi-private for-profit company very disappointing
  - See e.g. https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/

Q3: LLMs and Paywalls

- Some groups (e.g. EleutherAI) have replicated GPT3 and made it public:
  - https://github.com/EleutherAI/gpt-neo/
  - Still: duplicating huge amount of computational work is disappointing
  - A “CERN-like” effort for LMs: https://bigscience.huggingface.co
- Future of large LLMs:
  - I don’t think most will be paywalled, but that’s just a conjecture
  - I think the pendulum is swinging away from ever-larger models trained on ever-more data
  - Towards more data-efficient models that generalize more robustly
    - Bender and Koller 2020, Bender Gebru McMillan-Major Shmitchell 2021, Linzen 2020

Q4: Where to go next?

- “Could you offer some comment as to what the successful student would be able to do after completing the course vs. what people are doing in Deep Learning generally in the field? For example, after 572, I felt like I had a decent grounding in the fundamentals and that I could read (a great deal of) the literature (but not all of it.). I had the basics enough that I could read the scikit-learn documentation and build and evaluate models. But I was only familiar with a subset of the possible learners, and I was still floundering about a little until I had access to experts again. What would you expect an outcome of 575K to be along those lines? How much do we know vs. how much we should know? The course is only 10 weeks; what do you wish you could have included if you had more time? How would the interested student go about learning more? What are the cool kids doing these days?”

Q4: Where to go next?

- Outcome of 575K:
  - Ability to read and digest current state-of-the-art NLP papers
  - “ELMo is a bidirectional LSTM language model”
  - “BERT = Bidirectional Encoder Representations from Transformers”
  - Awareness of all of the key components of training such systems, so that you can do a deeper dive on specifics that interest you
  - Most importantly: deeper theoretical understanding of the models and their assumptions, so that you know what’s happening under the hood

Q4: Where to go next?

- What would I have included if we had more than 10 weeks?
  - More detail on practical considerations when training NLP models
  - More tasks: parsing, question-answering, toxic language detection, …
  - One complete project, with less scaffolding
    - But, e.g., that's what 573 and various 575s are for

Q4: Where to go next?

- Where to learn more?
  - Read papers and chase references when confused
  - Cornell's course has lots of online materials: http://www.phontron.com/class/nn4nlp2021/
  - Stanford CS224U (pre-recorded videos) http://web.stanford.edu/class/cs224u/
    - And CS224N (live lectures) http://web.stanford.edu/class/cs224n/
- NB: more about all of this on very last day of the course as well :)

Q4: Where to go next?

- “What are the cool kids doing these days?”
  - I’d like to reject the framing; there are no cool kids, and our goal as researchers should not be to simply chase trends
  - Over time, you’ll gain more and more familiarity with and confidence in the field, and hone your own perspective on what’s good and valuable research
  - Trust yourself, and pursue work that seems meaningful, useful, valuable, etc.

Thanks! Open floor for general discussion.