Modeling a Messy Language
"Araby"
Rami Botros, ramibotros@gmail.com
Video presentation: https://youtu.be/tb6ox_nKlYs

Introduction
- We build an RNN language model for Araby.
- Araby is a romanized chat language representation used by many Arabs online.
- It works by transliterating Arabic words into English text that would sound similar.

Example:
Ana raye3 el gam3a el sa3a 3 el 3asr.

- Numerals like 2,3,5 & 7 are used to represent sounds that do not exist in English.

Useful because:
- Fully representable in ASCII.
- Some people can speak but not write Arabic.

A Messy Language
- Araby emerged organically from (early) internet users, still has no defined rules.
  - "Which English vowel would sound most right?"
- Mostly used for colloquial chat, which is itself messy.
- User mix in a lot of English/French phrases.
- Araby is highly morphological. No spaces here:
  
  matetebloosh
  Do not (you) write it for him

Task
Conventional Language Modeling Task
- On character tokens (because of morphology)
- Inputs: sequence of context characters
- Output: probability distribution of next character
- Loss function: cross-entropy, Metric: Bits/Char

Data
LDC2017T07 Dataset, chat & SMS data
- 184K sentences, 3.7M running characters
- 52 unique characters
- 10K sentences for dev, 10K for test set.

Network Overview

Results
- BPC: Train 0.567 | Dev 0.561 | Test 0.553

Search for Model / Hparams
- LSTM vs GRU → LSTM sig. better here
- RNN size → relatively unimportant
- Number of stacked RNN layers → 1 layer best
- Dropout on input/output layers → ineffective
- Batchnorm on input/output layers → works great!
- Embedding/Output weight tying → works with tweak
- SGD vs Adagrad vs Adam → Adam sig. Better
- Gradient clipping (norm) → ~0.4 works best
- Early Stopping; rollback after 3 consec. non-improvements on Dev BPC

Conclusion / Outlook
- Model works but long / rare sequences still tricky.
- Text generation seems biased towards same head patterns.
- Definitely needs more data.

References
- Zhou Bang, BOLT: Egyptian-Arabic SMS/Chat and Transliteration LDC2017T07
- Martin Abadi et al, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems
- Paszke et al., "Automatic differentiation in PyTorch"
- Stephen Merity, Noah Shazeer, and Richard Socher, "Regularizing and optimizing LSTM language models"
- Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift"