CS 224n Assignment 4.

Abhishek Goswami.

February 6, 2019

1. Neural Machine Translation with RNNs

   (a)

   (b)

   (c)

   (d)

   (e)

   (f)

   (g) enc_masks
   - Using enc_masks we end up setting e_t to -inf where enc_masks has 1
   - This is required for the attention computation.
   - For attention computation, we want to compute the probability distribution over the words in the sentence
   - We do not want to include the padded words (that was only an implementation detail)
   - By using enc_masks, we make sure we are computing the softmax for the words in the original sentence

   (h)

   (i) Corpus BLEU: 22.708192645431552

   (j) Attention mechanisms

      i. Dot product attention
      - Advantage : Simpler model. Fewer number of parameters.
      - Disadvantage : Does not learn from the encoder hidden state

      ii. Multiplicative attention
      - Advantage : Tries to learn from the encoder hidden state, by doing a linear projection W_attProj over the hidden states in the encoder.
      - Disadvantage : More parameters. May lead to overfitting.

      iii. Additive attention
      - Advantage : Tries to learn from the encoder hidden state and also the decoder hidden state.
      - Disadvantage : Even more parameters. Prone to overfitting.

2. Analyzing NMT Systems

   (a) Understanding NMT errors

      i. • Error : *favorite of my favorites.* Does not make sense.

- Reason : The system is probably getting confused by reference to "one" favorite. So the NMT system says "another favorite" followed by "of my favorites" which does not make sense.
- Fix Suggestion : Add training data for a phrase like "another favorite of mine"

ii.
- Error : *author for children*. Incorrect meaning.
- Reason : Probably caused by overfitting caused by Multiplicative attention. The NMT system seems to be focussing on "ninos" (children) and "escribir" (write) . So it spits out "author for children" just fine. But lost context of being "widely read".
- Fix Suggestion : Try dot product attention to reduce overfitting.

iii.
- Error : *unk* word
- Reason : Word not present in the vocabulary
- Fix Suggestion : Use word segmentation, character-based models or hybrid NMT

iv.
- Error : *go back to the apple*
- Reason : Error because of linguistic construct. "Apple" literally translates to "manzana", however "around the block" translates to "alrededor de la manzana" . Too much attention paid to word manzana.
- Fix Suggestion : Combination of reduce overfitting (dot product attention) + more training data for colloquial words like "manzana"

v.
- Error : *women's room*
- Reason : Based on the context ("She") the NMT system used "women", instead of "teachers". Represents some bias in training data.
- Fix Suggestion : Fix bias in training data especially related to word "professores"

vi.
- Error : *100,000 acres* is incorrect quantification for *100,000 hectareas*
- Reason : 1 hectareas is not same as 1 acres
- Fix Suggestion : When we identify a word as a unit of measure, maybe it is best to keep the word as is in the translation.

(b) Explore outputs

i.
- Source Sentence : Si no son las vacunas, qu es?
- Reference Translation : If it isn't vaccines, what is it?
- NMT Translation : If you're not vaccines, what is it?
- Error : Phrase incorrect "you're not" instead of "it isn't"
- Reason : Seems to be a problem with sequence model decoders
- Fix Suggestion : Maybe a complex model e.g. Additive attention would have helped.

ii.
- Source Sentence : Yo estaba asombrada.
- Reference Translation : I was in awe.
- NMT Translation : I was [unk]
- Error : Unknown word.
- Reason : Rare and unknown word problem.
- Fix Suggestion : Use word segmentation, character-based models or hybrid NMT

(c) BLEU scores

i. For c1 :

- c : 5
- r* : 4
- BP : 1
- p1 : 3/5
- p2 : 2/4
- BLEU : 0.5477225575051662. [np.exp(0.5*np.log(3/5) + 0.5*np.log(2/4))]

For c2 :

- c : 5
- r* : 4
- BP : 1
- p1 : 4/5
- p2 : 2/4
- BLEU : 0.6324555320336759. [np.exp(0.5*np.log(4/5) + 0.5*np.log(2/4))]

As per the BLEU scores, c2 is a better translation. I agree. c2 makes more sense compared to c1.

ii. For c1 :

- c : 5
- r* : 6
- BP : exp(-1/5)
- p1 : 3/5
- p2 : 2/4
- BLEU : 0.4484373019840029. [np.exp(-1/5) * np.exp(0.5*np.log(3/5) + 0.5*np.log(2/4))]

For c2 :

- c : 5
- r* : 6
- BP : exp(-1/5)
- p1 : 2/5
- p2 : 1/4
- BLEU : 0.2589053970151336. [np.exp(-1/5) * np.exp(0.5*np.log(2/5) + 0.5*np.log(1/4))]

As per the BLEU scores, c1 is a better translation. I disagree. To me it seems c2 is a better translation.

iii. The example above shows that evaluation with respect to a single reference can be problematic.

- A particular source sentence can be expressed in several different ways in the target language.
- It makes sense to have several different reference translations.
- The robust score of a NMT translation must be based on several different reference translations, not just one reference translation.

iv. Advantages of BLEU compared to human evaluation.

- Deterministic BLEU scores. Any score based on human evaluation will tend to vary (based on human judges)
- Ability to scale the evaluation process to millions of translations. Human evaluation does not scale.

Disadvantages of BLEU.

- A good translation may get low BLEU score since BLEU scores are based on n-grams. Human evaluation goes much beyond n-gram counts.
- no ability to penalize for offensive / bad words etc. Human evaluation is suitable for penalizing presence of profanity in translations etc.