Improved Beam Search Diversity for Neural Machine Translation with k-DPP Sampling

Jon Braatz, Max Spero

Abstract

Beam search is widely used in natural language generation tasks to evaluate decoder outputs and translate them into comprehensible sentences. However, the usual method of search by choosing the top-k most probable extensions could lead to the beam getting populated with correlated samples which reduces efficiency and could crowd out other viable outputs. We explore an update to beam search, by sampling using determinantal point processes. This promotes diversity of samples evaluated and mitigates correlations, resulting in a search over a more diverse sample space. We find that a hybrid approach is able to match the BLEU score of top-*k* approaches while placing a focus on hypothesis diversity.

Beam search algorithms

Algorithm 1 Top-k beam search

Input: Vocabulary V, beam size k

$H_0 \leftarrow \{[< s >]\}$
$H_{complete} \leftarrow \emptyset$
$t \leftarrow 0$
while $|H_{complete}| \neq k$ **do**
  $k_{live} \leftarrow k - |H_{complete}|$
  $H_{temp} \leftarrow$ top_scores$(k_{live}, \{[h, w] | h \in H_t, w \in V\})$
  $H_{complete} \leftarrow \{h | h \in H_{temp}, h_{t+1} =< \backslash s >\}$
  $H_{t+1} \leftarrow \{h | h \in H_{temp}, h_{t+1} \neq< \backslash s >\}$
  $t \leftarrow t + 1$

Algorithm 2 k-DPP beam search

Input: Vocabulary V, beam size k, pool size αk

$H_0 \leftarrow \{[< s >]\}$
$H_{complete} \leftarrow \emptyset$
$t \leftarrow 0$
while $|H_{complete}| \neq k$ **do**
  $k_{live} \leftarrow k - |H_{complete}|$
  $H_{pool} \leftarrow$ top_k$(S, \{[h, w] | h \in H_t, w \in V\})$
  $H_{temp} \leftarrow$ kdpp_sample$(k_{live}, H_{pool}))$
  $H_{complete} \leftarrow \{h | h \in H_{temp}, h_{t+1} =< \backslash s >\}$
  $H_{t+1} \leftarrow \{h | h \in H_{temp}, h_{t+1} \neq< \backslash s >\}$
  $t \leftarrow t + 1$

Algorithm 3 Hybrid beam search

Input: Vocabulary V, DPP-beam size k, pool size αk, top-beam size m

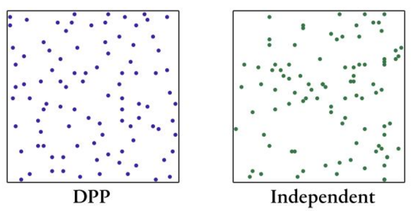
$H_0 \leftarrow \{[< s >]\}$
$H_{complete} \leftarrow \emptyset$
$t \leftarrow 0$
while $|H_{complete}| \neq k$ **do**
  $k_{live} \leftarrow k - |H_{complete}|$
  $H_{temp} \leftarrow H_{kDPP}(V, k, S, H_t) \bigcup H_{topK}(V, m, H_t)$
  $H_{complete} \leftarrow \{h | h \in H_{temp}, h_{t+1} =< \backslash s >\}$
  $H_{t+1} \leftarrow \{h | h \in H_{temp}, h_{t+1} \neq< \backslash s >\}$
  $t \leftarrow t + 1$

Determinantal Point Processes

- Each step of beam search gives us a set of candidate partial translations and their scores. Want to choose k of them with scores balanced with diversity.
- DPP is probability distribution over subsets. k-DPP conditions on cardinality k
- For candidate feature unit vectors φ_i and φ_i with associated "quality scores" q, score subset Y according to:

$$S_{ij} \equiv \phi_i^\top \phi_j \qquad \mathcal{P}_L(Y) \propto \left(\prod_{i \in Y} q_i^2\right) \det(S_Y),$$

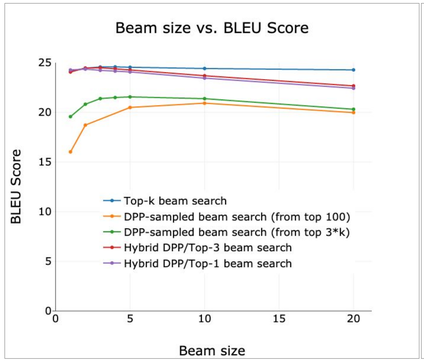
- We set q_i = sqrt(score)$_i$ and use decoder hidden state as feature vector for candidates
- Sample subset of k new candidates from distribution

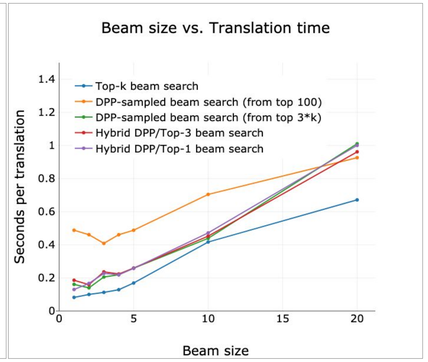


Conclusions

- We found that DPPs had both the theoretical backing and real-world potential to improve diversity of beams within beam search.
- Real-world analysis showed that pure DPP sampling underperformed top-*k*, but a hybrid approach seemed to track top-*k* performance.
- Qualitatively, increased beam sample diversity was observed when using DPP methods.
- When DPP sampling and top-*k* are combined we see potential for a powerful situational tool for translating complex sentences.

Analysis





Plots of how different beam search algorithms performed with different beam sizes.
(Left) plots beam size against a test set BLEU score.

- We observe that BLEU score seems to peak around *k*=5 and decrease from there.
- We also observe that the top predicted hypothesis seems to be by far the most important for a good translation.

(Right) plots beam size against translation time.

- While DPP and hybrid approaches take longer than top-*k* beam search, these results show that beam size is still the largest factor in determining how long a translation takes.

Source sentence: Y me dije: "Un momento, esto parece interesante."
Reference translation: And I was like, "Hold on. That sounds interesting."
Top-*k* translation: And I said, "A moment, this seems interesting.
DPP translation: And I said, "Wait a moment, this seems interesting.

Top-*k*-sampled Beam	***k*-DPP-sampled Beam**
`['And', 'I', 'said', 'to']`	`['And', 'I', 'said', 'to']`
`['And', 'I', 'said,', '"A']`	`['And', 'I', 'said,', '"A']`
`['And', 'I', 'said,', '"One']`	`['And', 'I', 'said,', '"One']`
`['And', 'I', 'said,', '"It']`	`['And', 'I', 'said,', '"Wait']`
`['And', 'I', 'said,', '"It\'s']`	`['And', 'I', 'said,', '"You']`

Runtime Analysis: We determined that for a vocabulary of size $|V|$, beam size of k, and maximum sentence size of m, pre-sampling (before DPP) from a set of size S:

Top-*k* beam search = $O(|V|k^m)$
***k*-DPP beam search =** $O(|V|S^3k^m)$
But by setting S to be a factor of k, we can reduce it to simply
***k*-DPP beam search =** $O(|V|k^m)$

References

Ashwin K. Vijayakumar et al. "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models". In: CoRR abs/1610.02424 (2016). arXiv: 1610.02424. URL: http://arxiv.org/abs/1610.02424.

Alex Kulesza, Ben Taskar, et al. "Determinantal point processes for machine learning". In: Foundations and Trends in Machine Learning 5.2–3 (2012), pp. 123–286.