Mini-batch Parallelized Beam Search

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Motivations

- 우리는 beam search를 통해 greedy search의 성능을 개선할 수 있다.

- 하지만 beam search를 주어진 mini-batch에 대해서
  각각 수행하는 것은 매우 비효율적인 작업
  - 따라서 mini-batch parallelized beam search를 구현하면 속도와 성능 둘 다 잡을 수 있다!

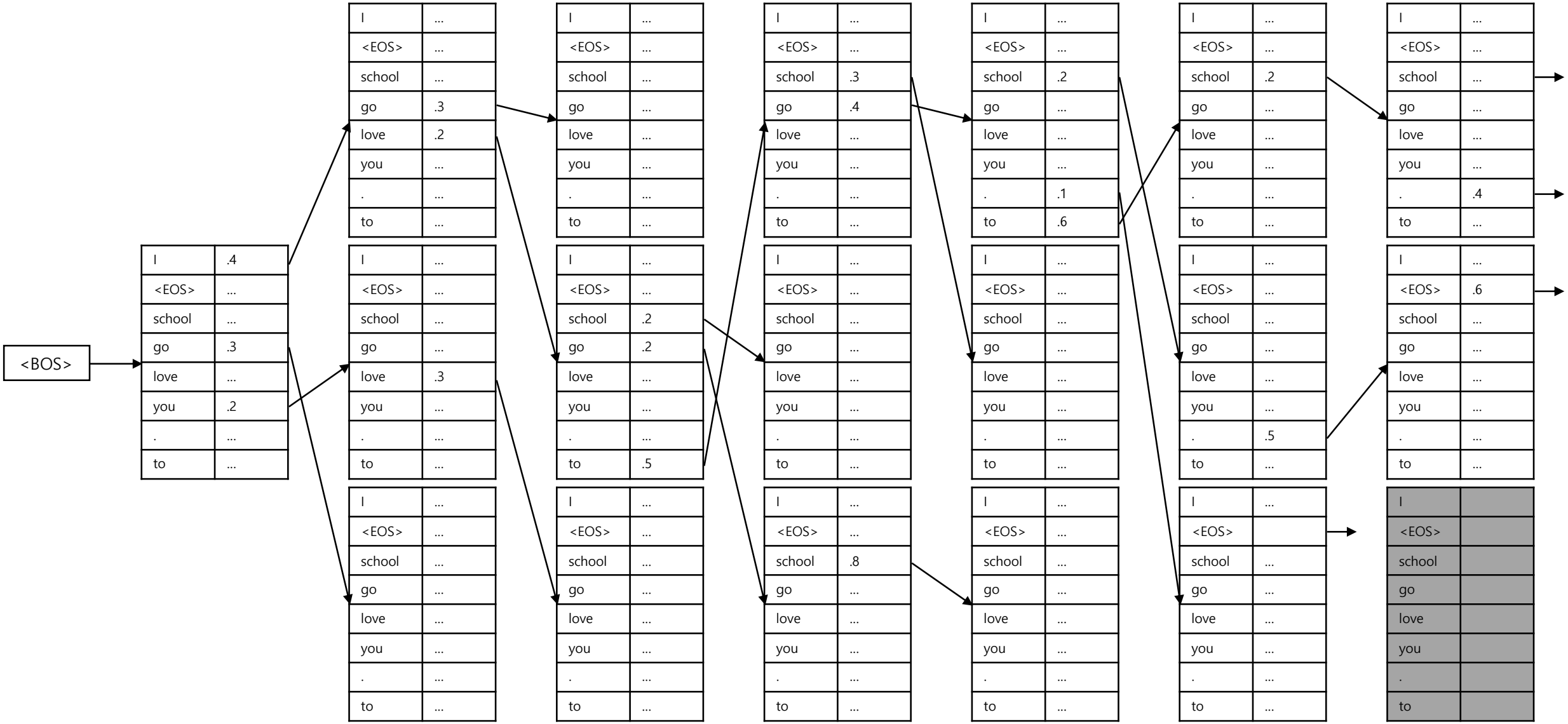
Solution

- 한 샘플에 대한 beam search를 수행할 때, k번의 inference가 수행됨
  - 마치 k개의 샘플에 대한 inference로 볼 수 있음

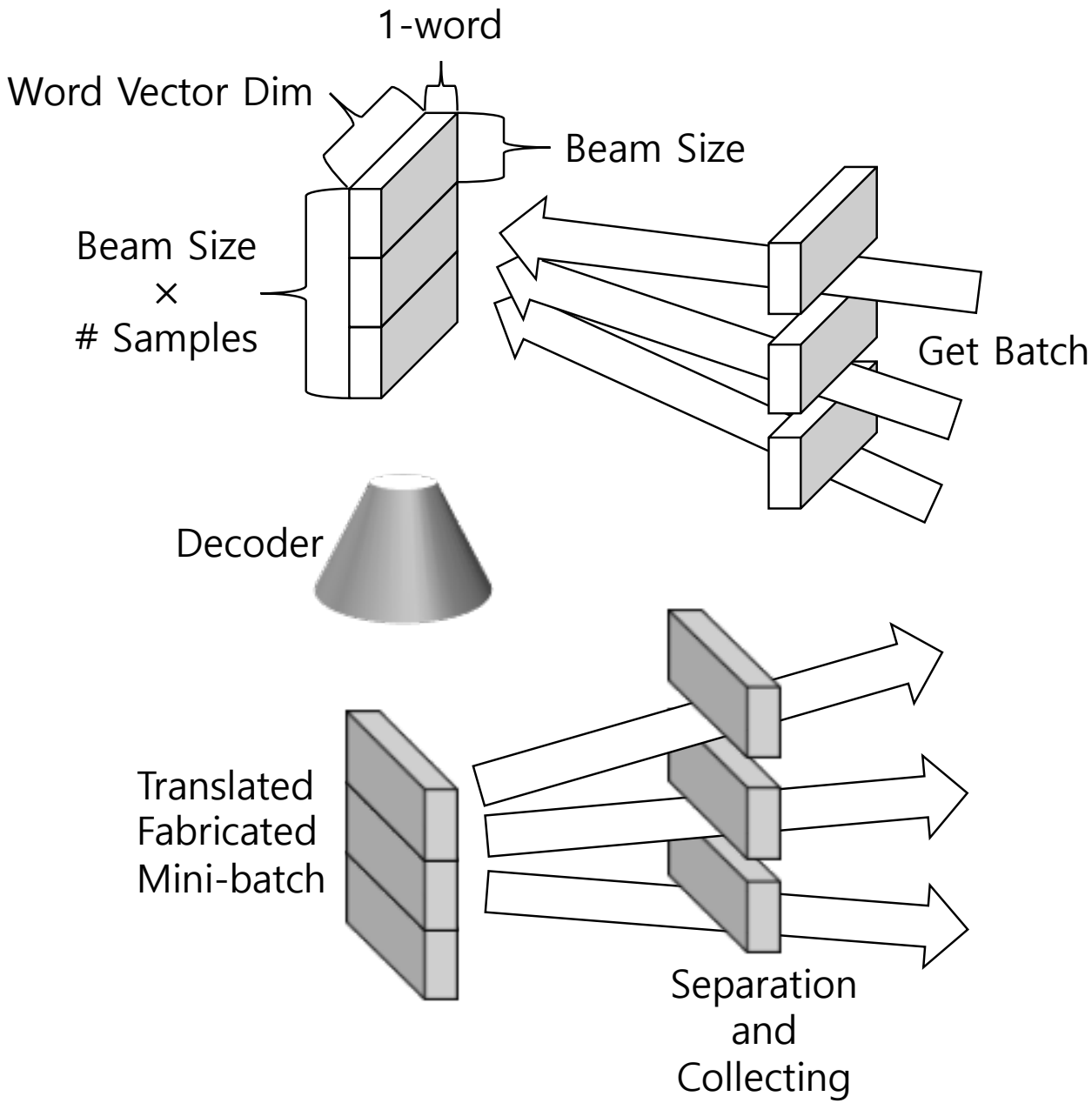
- n 샘플에 대한 beam search를 수행할 때, $n \times k$번의 inference가 수행됨
  - 마치 $n \times k$개의 샘플에 대한 inference로 볼 수 있음

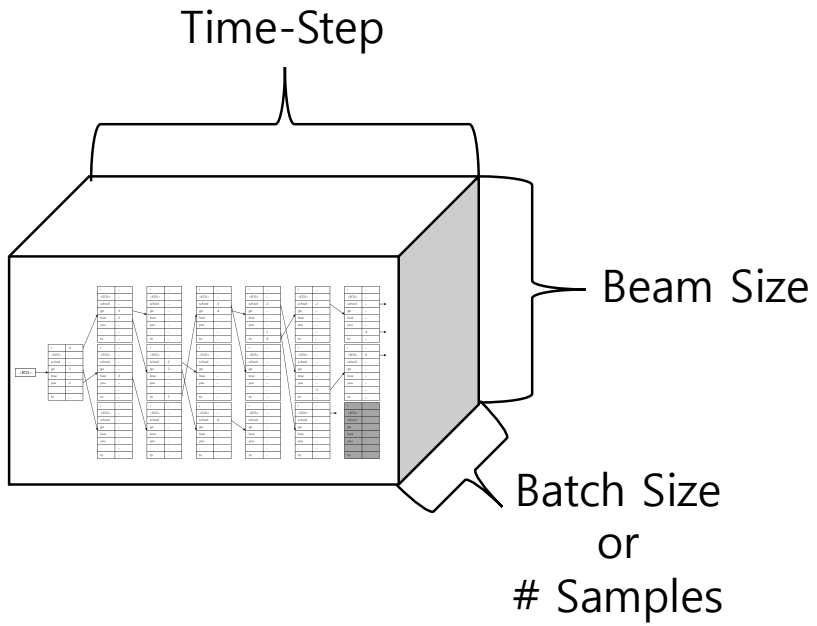
- 따라서, $n \times k$개 샘플의 mini-batch에 대한 inference를 수행하면 됨

Parallelized Beam Search (for 1 sample)



Decoding Process Overview





Conclusion

- Mini-batch parallelized beam search를 통해 성능과 속도 모두 잡을 수 있음
  - 하지만 속도가 무작정 빨라지는 것은 아님

- 실제 deploy 환경에서는 beam 갯수 조절을 통해
  속도와 성능 사이의 trade-off를 조절
  - 실제 배포를 위해서는 code profiling을 통해 속도 저하를 발생시키는 코드를 제거해야 함