Neural Language Model

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Neural Language Model

- Resolve Sparsity
  - Training set
    - 고양이는 좋은 반려동물 입니다.
  - Test set
    - 강아지는 훌륭한 애완동물 입니다.

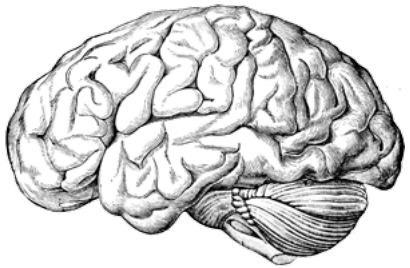
Unseen Word Sequence

Neural Language Model

- Resolve Sparsity
  - Training set
    - 고양이는 좋은 반려동물 입니다.
  - Test set
    - 강아지는 훌륭한 애완동물 입니다.

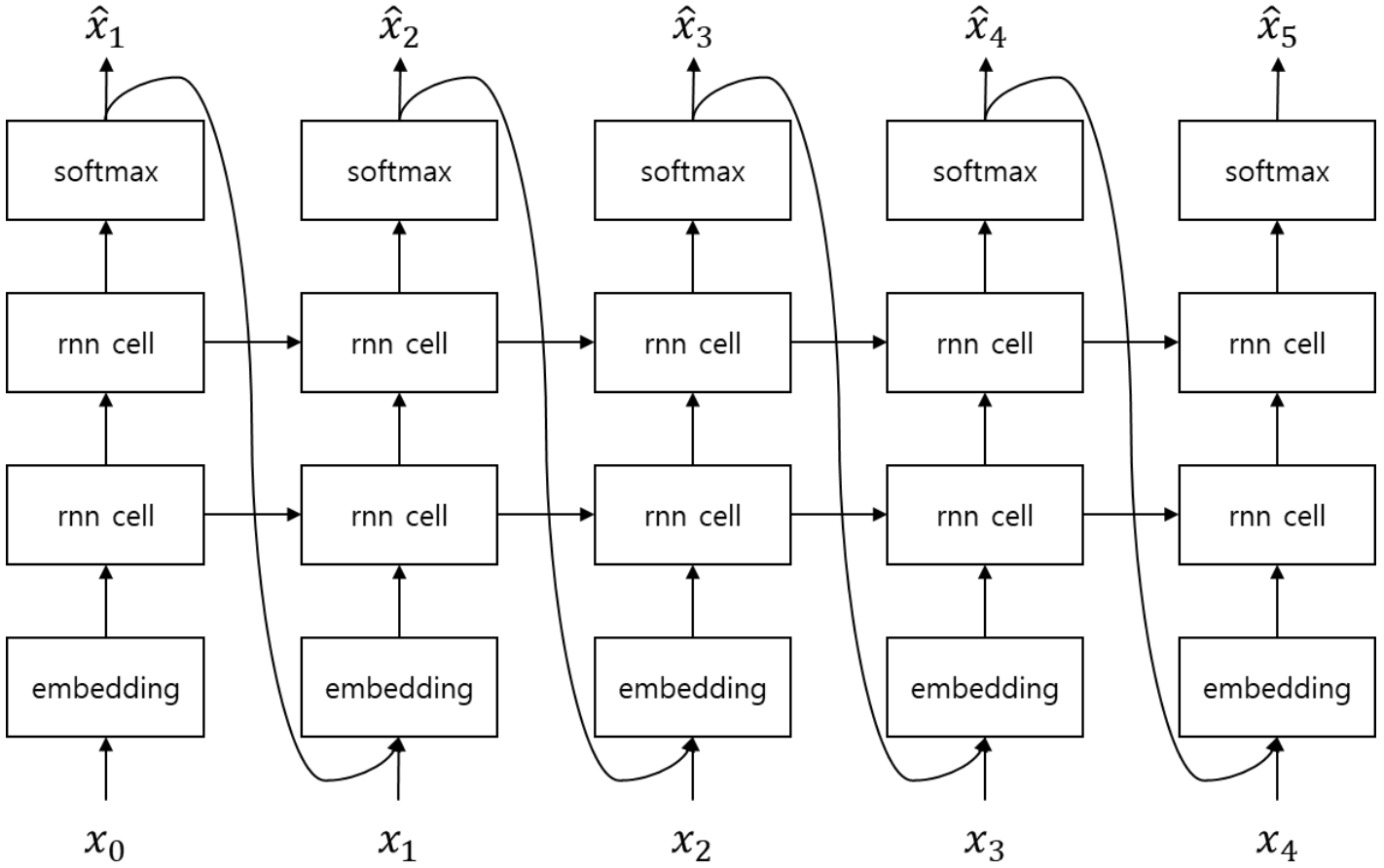
Unseen Word Sequence

- Because we know (and we can approximate that)
  - 고양이 ≈ 강아지
  - 좋은 ≈ 훌륭한
  - 반려동물 ≈ 애완동물



- But n-gram CANNOT, because words are discrete symbols.

Neural Language Model



Neural Language Model

- Find parameter that maximize likelihood for given training corpus.

$$\mathcal{D} = \{x^i\}_{i=1}^N$$

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(x^i; \theta)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} \log P(x_j^i | x_{<j}^i; \theta)$$

Neural Language Model

- Take a step of gradient descent to minimize negative log-likelihood.

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{n} \log P(x_j^i | x_{<j}^i; \theta)$$

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$$

$$\log P(x_t | x_{<t}; \theta) = x_t^T \cdot \log f_\theta(x_{t-1}, h_{t-1}),$$

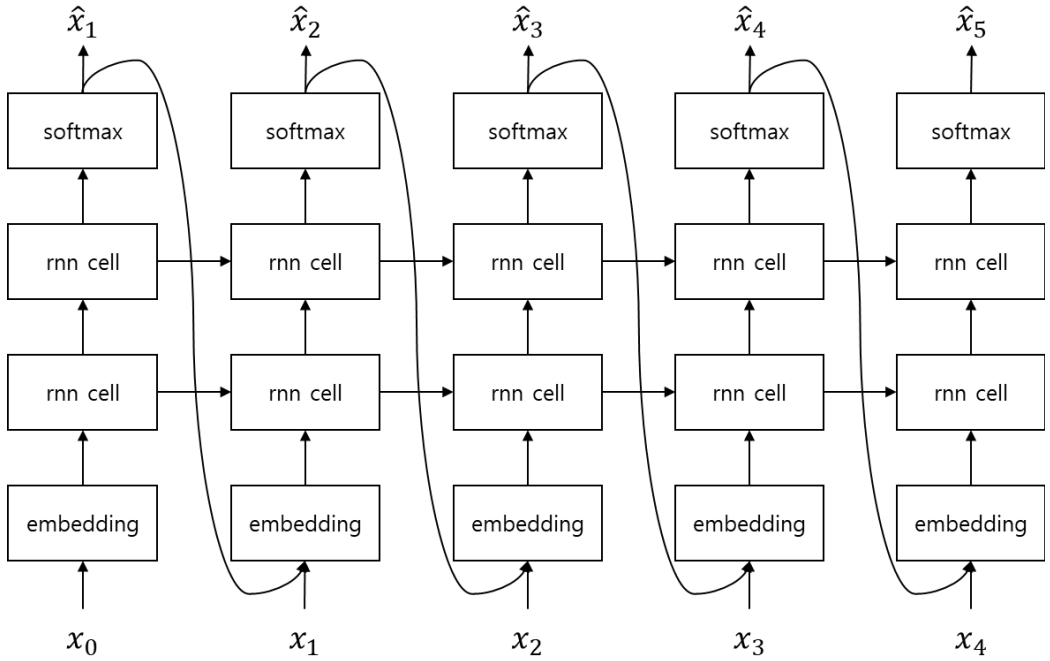
where x_t is one-hot vector, and f_θ is model with parameter θ.

Neural Language Model

- Detail:

$$\log P(x_t|x_{<t};\theta) = x_t^T \cdot \log f_\theta(x_{t-1}, h_{t-1}),$$
where x_t is one-hot vector, and f_θ is model with parameter θ.

$$\begin{aligned} f(x_{t-1}, h_{t-1}) &= \text{softmax}(\text{RNN}(\text{emb}(x_{t-1}), h_{t-1}) \cdot W), \text{ where } W \in \mathbb{R}^{\text{hidden\_size} \times |V|} \\ &= \text{softmax}(h_t \cdot W), \text{ where } h_t \in \mathbb{R}^{\text{batch\_size} \times \text{hidden\_size}} \\ &= \hat{x}_t \end{aligned}$$
where $\hat{x}_t$ is a probability distribution that $P(\cdot|x_{<t};\theta)$.



Loss Function of NNLM

- Find theta that minimize negative log-likelihood.

- Find theta that minimize cross entropy with ground-truth probability distribution.

Cross Entropy Loss

Summary

n-gram (previous method)

- 단어를 discrete symbol로 취급
  - Exact matching에 대해서만 count
- 따라서 generalization issue 발생
  - Markov Assumption 도입 (n-gram)
  - Smoothing & Discounting
  - Interpolation & Back-off
  - Unseen sequence에 대한 대처 미흡
- 빠른 연산 & 쉽고 직관적
  - 단순한 look-up table 방식
  - 문장 fluency 비교 task에서는 괜찮음

Neural Network Language Model

- Word embedding을 통해, unseen sequence에 대해 대처 가능
- Generation task에서 특히 강점
- 연산량 많음 (feed forward 연산)
  - 해석(XAI) 난이도 증가