Hito 1
Language Modeling

Pablo Cleveland
Camilo Escobar
Diego Garrido
Pablo Miranda

10 de Septiembre de 2019

Agenda

1. Motivación
2. Task
3. Dataset
4. Métricas
5. Modelo y Baseline

1

Motivación

Aplicaciones

Machine Translation

Speech Recognition

Optical Character Recognition (OCR)

2

Task

¿Qué es *Language Modeling?*

“El objetivo del modelamiento de lenguaje estadístico es aprender la función de probabilidad conjunta de secuencias de palabras en un lenguaje”*

* A Neural Probabilistic Language Model, Yoshua Bengio et al.

Función de Probabilidad Conjunta

$$P(w_0, \dots, w_N) = P(w_0) \prod_{i=1}^{N} P(w_1 \mid w_0, \dots w_{i-1})$$

w_i = palabra del vocabulario

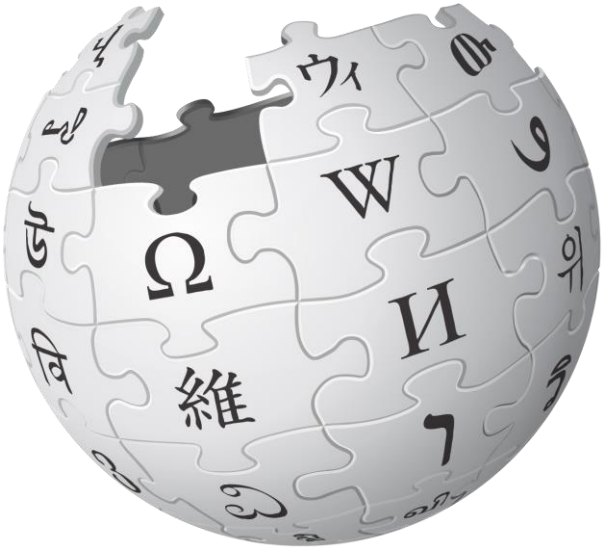
Input: Secuencia ordenada de palabras.
Output: Vector de probabilidades sobre el vocabulario.

3

Dataset

WikiText-103

- Construido a partir de artículos de Wikipedia
- 103 millones de *Tokens*
- 260.000 palabras



	Train	Valid	Test
Artículos	**~20.5 M**	**60**	**60**
Tokens	**~103MM**	**~218M**	**~245M**

Tabla 1: separación conjuntos de datos.

4

Métricas

Perplexity

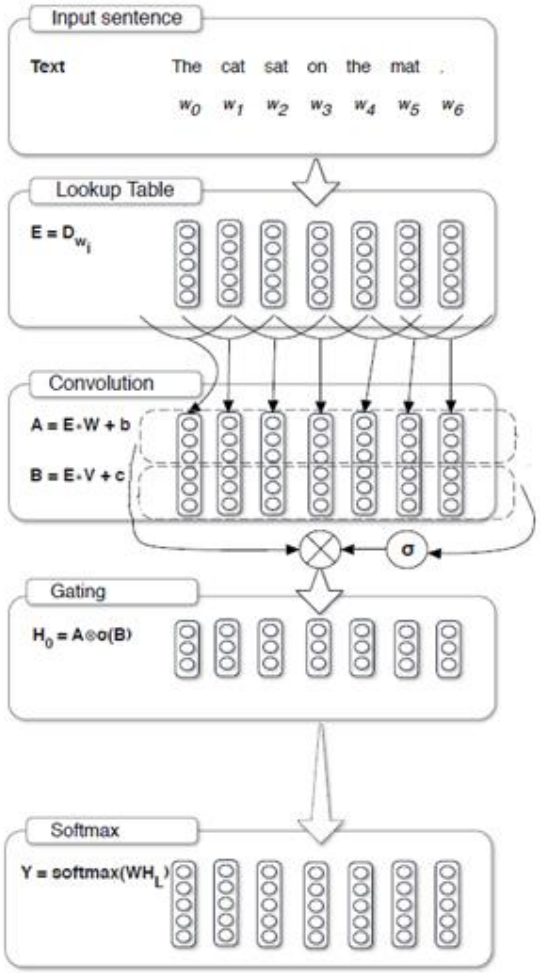
- Medida de cuán bien un modelo probabilístico predice.
- Se mide sobre una muestra de test.
- Mientras menor mejor, dominio práctico [1, |V|], donde V es el vocabulario.

$$\exp\left(\frac{1}{N}\sum_{i=1}^{N} -ln(p(w_i|\ldots,w_{i-1}))\right)$$

5

Modelo y Baseline

Gated Convolutional Neural Network (GCNN)



Baseline

Modelo de Lenguaje de Trigramas con Interpolación Lineal

$$p(w_0, \ldots, w_N) = \prod_{i=0}^{N} (\lambda_1 \times q(w_i|w_{i-2}, w_{i-1}) + \lambda_2 \times q(w_i|w_{i-1}) + \lambda_3 \times q(w_i))$$

Donde

$$\lambda_i \geq 0 \text{ y } \lambda_1 + \lambda_2 + \lambda_3 = 1$$

Referencias

- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155.
- Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In Proceedings of the 34th International Conference on Machine Learning- Volume 70, pages 933–941. JMLR. org.
- Edouard Grave, Armand Joulin, Moustapha Cisse, Herve Jegou, et al. 2017. Efficient softmax approximation for gpus. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 , pages 1302–1310. JMLR. org
- Stephen Merity, Caiming Xiong, James Broadbury, and Richard Socher, 2016. Pointer Sentinel Mixture Models, arXiv e-prints, page arXiv:1609.07843
- Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759.

Gracias!

Hito 1
Language Modeling

Pablo Cleveland
Camilo Escobar
Diego Garrido
Pablo Miranda

10 de Septiembre de 2019