Adversarial Training with Contrastive Learning in NLP

Procesamiento de Lenguaje Natural

Francisco Javier Sáez Maldonado

13 de marzo de 2022

Máster en Ciencia de Datos

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Índice

Introducción

- **Tarea**: Modelado del lenguage (LM) y Traducción automática (NMT)

Introducción

- **Tarea**: Modelado del lenguage (LM) y Traducción automática (NMT)
- **Objetivo**: Conseguir modelos que sean más robustos semánticamente:

$$\text{Inputs parecidos} \implies \text{Outputs parecidos}$$

Herramientas

Herramientas

Adversarial Training

Adversarial Training

Definición (Adversarial Learning)

Técnica usada en el aprendizaje automático para, usando información sobre un modelo, crear ataques maliciosos para causar fallos en el modelo

Adversarial Training

Definición (Adversarial Learning)

Técnica usada en el aprendizaje automático para, usando información sobre un modelo, crear ataques maliciosos para causar fallos en el modelo

Definición (Adversarial example)

Ejemplo diseñado para engañar al modelo, creado introduciendo una perturbación en un ejemplo original.

¿ Cómo ayuda el aprendizaje adversario a nuestros modelos ?

Ejemplos adversarios en NLP

Ejemplos de técnicas:

- Visuales
- omega

Ejemplos adversarios

Dada una secuencia $s = \{x_1, \ldots, x_T\}$ de tokens

1. Creamos una representación embebida en un espacio continuo

$$\mathbf{E}x_i = e_i.$$

2. Añadimos una pequeña perturbación en el embedding

$$e_i' = e_i - \epsilon \frac{g}{\|g\|_2},$$

siendo $g = \nabla_{e_i} J(s, \theta)$ y J la función de coste.

Función de coste actual:

$$\mathcal{J}(\theta) = \sum_s \mathcal{L}(s, \theta) + \alpha \sum_{s'} \mathcal{L}_{adv}(s', \theta), \quad \alpha \in [0, 1].$$

Herramientas

Contrastive Learning

Contrastive Learning

Idea: Acercar las representaciones de ejemplos positivos (de la misma clase) y alejar las de los ejemplos negativos (resto de ejemplos).

Ejemplo

Original: Elefante. Positivo: Hipopótamo. Negativo: Pistola.

Definición (Pérdida contrastiva)

Sean a_i las entradas originales, p_{a_i} ejemplos positivos y n_a ejemplos negativos. Se define la pérdida contrastiva como:

$$\mathcal{L}_{cont} = -\sum_{a_i \in A} \log \frac{\exp(a_i \cdot p_{a_i}/\tau)}{\sum_{n_a \in A - \{a_i\}} \exp(a_i \cdot n_a/\tau)}$$

Framework

Resultados

Conclusiones

- El uso de la teoría de la información proporciona un buen punto de partida para el aprendizaje de representaciones.
- El aprendizaje contrastivo ha probado ser la mejor forma de obtener representaciones que son útiles en tareas posteriores.
- Ambos marcos de trabajo probados obtienen buenos resultados en la adaptación a conjuntos de datos más pequeños.

Gracias por su atención