Universidad Alfonso X el Sabio

Ejercicios de Programación Lineal

Germán Llorente Muñoz

October 9, 2023

October 9, 2023

1 Introducción

Los modelos de lenguaje grande han revolucionado la generación de texto, pero la manipulación adecuada de los hiperparámetros es fundamental para obtener resultados óptimos. En este paper, exploramos cómo la temperatura, num_beams, top_k y top_p influyen en las estrategias de decodificación y cómo ajustar estos hiperparámetros puede mejorar la calidad del texto generado.

2 Hiperparámetros en Estrategias de Decodificación

2.1 Temperatura

La temperatura controla la "creatividad" de las predicciones del modelo. Una temperatura baja favorece las opciones más probables, mientras que una temperatura alta introduce aleatoriedad. Por ejemplo, al utilizar una temperatura de 0.7 en ChatGPT, se pueden obtener respuestas más diversas y creativas.

2.2 num_beams

El num_beams determina cuántas secuencias alternativas el modelo debe considerar durante la búsqueda por beams. Si establecemos num_beams en 5, el modelo considerará 5 posibles continuaciones para cada paso, lo que puede mejorar la diversidad del texto generado. En la tarea a realizar, el número de beans está en 5, lo que hace que el código te prolongue la frase con las 5 palabras (o símbolos, los puntos y las comas cuentan como palabras) más probables teniendo en cuenta la frase que se ha escrito.

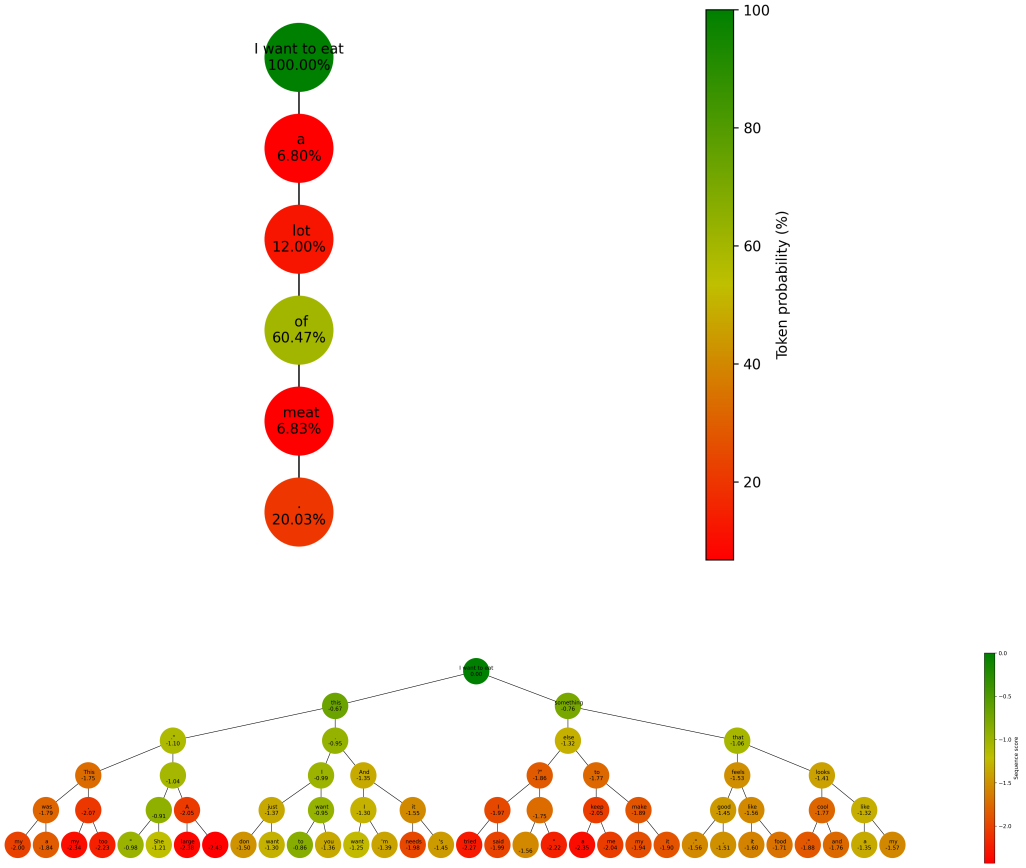
2.3 top_k y top_p

Estos hiperparámetros controlan el muestreo de tokens durante la generación. Si configuramos top_k en 20 y top_p en 0.85, el modelo solo considerará los 20 tokens más probables y aquellos cuya probabilidad acumulada sea al menos del 85%. Esto puede producir respuestas más coherentes y significativas. En el proyecto que hemos realizado, se veía esto ejemplificado cuando aumentaba el

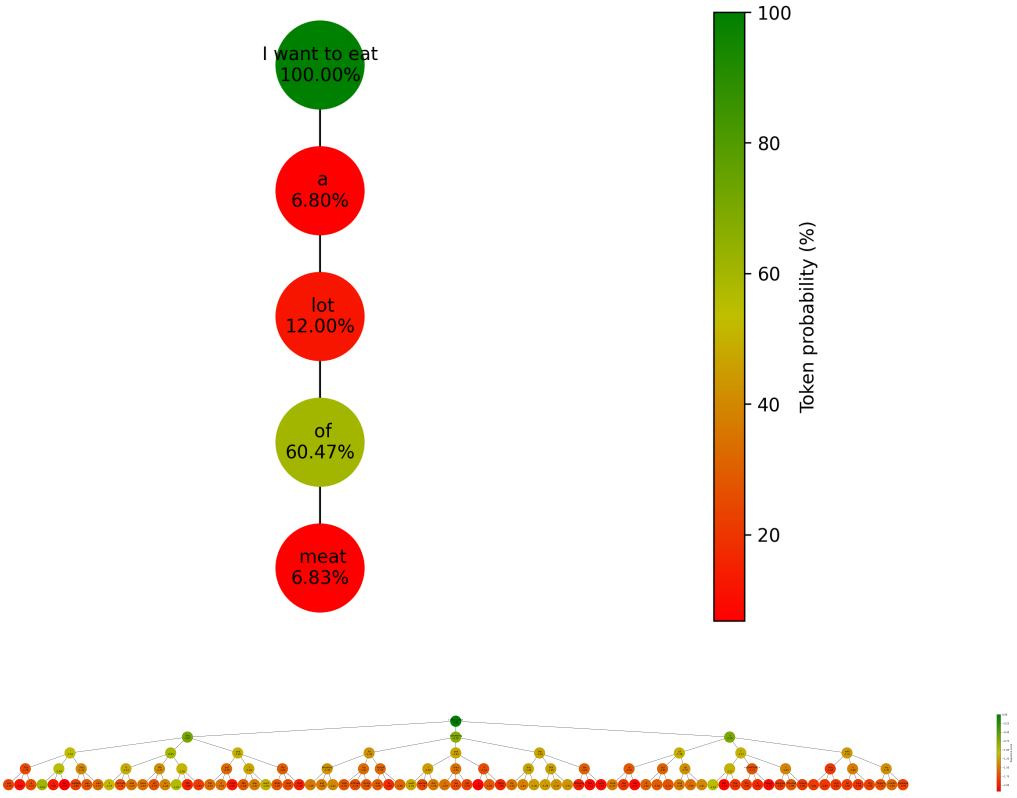
número de palabras a considerar cuando aumentabamos el top_k, pero las palabras que aparecían tenían una probabilidad tan baja que no eran interesantes de considerar.

3 Impacto en la Calidad del Texto Generado

3.1 Ejemplos de Manipulación



En este ejemplo hemos usado la frase inicial "I want to eat". Los beans son 2 y el top_k es 5. Esto se ve claramente en el árbol dado que su altura es 6 y cada elemento tiene dos hijos. Veamos que sucede si cambiamos estos parámetros.



Aquí hemos puesto beans 3 y top_k 4. Se han generado 4 palabras más allá de la oración inicial y se han considerado tres palabras por cada decisión.

4 Conclusiones

En este estudio, hemos explorado las complejidades de los modelos de lenguaje grande, centrándonos en la manipulación de hiperparámetros clave: temperatura, num_beams, top_k y top_p. Nuestro análisis detallado revela la influencia significativa que estos parámetros tienen en la generación de texto, destacando su impacto en la creatividad, diversidad y coherencia del texto producido. Los resultados siguientes se han obtenido al manipular el código fuente presente en el campus.

Al variar la temperatura, observamos cómo esta dimensión puede controlar el nivel de aleatoriedad en las respuestas generadas. Una temperatura más baja favorece las opciones más probables, mientras que una temperatura más alta introduce variabilidad y creatividad en el texto. Esto fue ejemplificado mediante el uso de una temperatura de 0.7 en ChatGPT, que resultó en respuestas más

diversas y creativas.

El hiperparámetro num_beams demostró ser crucial para la generación estructurada del texto. Al ajustar el número de beams, se pudo observar cómo la profundidad y diversidad del texto generativo variaban significativamente. Un num_beams más alto, como 5, lleva a una expansión más amplia de la oración inicial, mostrando la influencia directa en la longitud y complejidad de la respuesta generada.

La manipulación de top_k y top_p también reveló insights interesantes. Al limitar las opciones a los 20 tokens más probables y aquellos con una probabilidad acumulada del 85% o más, se produjeron respuestas más coherentes y significativas. Sin embargo, al aumentar top_k, aunque se consideraron más palabras, la baja probabilidad de algunas de estas opciones condujo a resultados menos interesantes.

Además, presentamos ejemplos visuales para ilustrar cómo la variación de los hiperparámetros num_beams y top_k afecta la estructura del árbol de decisión. Estos ejemplos proporcionan una comprensión intuitiva de cómo estos parámetros influyen en la generación del texto.

En resumen, este estudio resalta la importancia de la manipulación cuidadosa de los hiperparámetros para obtener resultados textuales deseados. Ya sea buscando respuestas creativas, coherentes o específicas, ajustar estos parámetros es esencial. Estos hallazgos no solo ofrecen una visión profunda de los modelos de lenguaje, sino que también proporcionan pautas prácticas para los desarrolladores que buscan mejorar la calidad y relevancia del texto generado en diversas aplicaciones.