Top K, Top P and Temperature in AI text generation

Introduction

Imagine you are an AI assistant in a large library. Your job is to choose books for visitors based on their requests. How you choose these books determines how varied and surprising your recommendations are. In the world of AI text generation, Top K, Top P and Temperature play a similar role. They determine how the AI chooses words when creating sentences.

In this guide, we will explore these three important concepts using an ongoing example. We will see how an AI model chooses the next word in the sentence ***"The weather is today ".***

Top K, Top P and Temperature in AI text generation

Origin of the names "Top K" and "Top P"

Top K

"Top K" stands for "Top K selection."

- The "K" in Top K is simply a variable indicating the number of top choices we want to keep.
- In mathematics and computer science, the letter "K" is often used as a general constant or parameter, especially when it comes to counting or ranking items.
- "Top" refers to the fact that we select the K most likely (or "top") options.

So "Top K" literally means "the top K items," where K is a number we can choose ourselves .

Top P

"Top P" stands for "Top Probability" or "Top Percentage."

- The "P" refers to the cumulative probability (Probability) or percentage (Percentage) that we use as a threshold.
- "Top" here indicates that we select the most likely options until we reach the P threshold.

Top P is also called "nucleus sampling" because it selects the "core" (nucleus) of the most likely options.

Comparison

- Top K selects a fixed number (K) of options, regardless of their probabilities.
- Top P selects a variable number of options, but with a fixed total probability (P).

Both methods are designed to retain the "top" or most relevant choices, but do so in different ways: Top K looks at numbers, while Top P looks at cumulative probabilities.

Top K, Top P and Temperature in AI text generation

What exactly do Top K, Top P and Temperature do?

Top K: The Limited Bookcase

Top K is like a librarian working with a fixed, limited bookcase.

- If K is set to 5:
  - The librarian places only the 5 most popular books on the shelf.
  - All choices are made from these 5 books, regardless of how popular they are relative to each other.

Top P: The Selective Bookcase

Top P, "nucleus sampling," works like a selective bookcase.

- Suppose Top P is set to 0.8 (80%):
  - The AI looks only at the most likely words that make up 80% of the total probability.
  - It is as if the librarian only looks at the most popular 80% of the books and ignores the rest.

Temperature: The Thermostat of Creativity

Imagine Temperature as a thermostat that controls the "temperature" of the AI's creativity.

- **Low Temperature** (close to 0): The AI is like a cautious librarian choosing only the most obvious books. Answers are predictable and consistent, but possibly less creative.
- **High Temperature** (direction 1 or higher): The AI becomes an adventurous librarian who also dares to recommend obscure or unexpected books. Answers are more varied and creative, but possibly less coherent.

Top K, Top P and Temperature in AI text generation

How do they work?

Let's use our continuous example to show how Top K, Top P and Temperature work together. The AI must choose the next word in the sentence:
"The weather today is...".

Top K in Action

1. The AI ranks all possible subsequent words by probability.
2. Only the top K most likely words are retained.
3. The AI chooses words from these K, weighted by their probability.

Example:
- K = 5
- All the possible words and their chances:
  1. "sunny" (30%)
  2. "rainy" (25%)
  3. "cloudy" (20%)
  4. "warm" (10%)
  5. "cold" (8%)
  6. "stormy" (5%)
  7. "foggy" (2%)

After Top K (K=5) remain:
- "sunny" (30%)
- "rainy" (25%)
- "cloudy" (20%)
- "warm" (10%)
- "cold" (8%)

Top P in Action

1. Of the remaining words after Top K, the AI adds up the probabilities.
2. Once the sum reaches the Top P value, all remaining words are excluded.

Example:
- Top P = 0.8 (80%)
- Remaining words after Top K: "sunny"
  (30%)
  "rainy" (25%)
  "cloudy" (20%)
  "warm" (10%)
  "cold" (8%)

After Top P (P=0.8) remain:
- "sunny" (30%)
- "rainy" (25%)
- "cloudy" (20%)
(These add up to 75%, which is the closest to 80% without going over it)

Temperature in Action

1. Temperature adjusts the probabilities of the remaining words.
2. Lower Temperature increases the differences, higher Temperature decreases them.

Example with remaining words after Top P:
- At low Temperature (0.3):
 "sunny" (greatly increased probability, e.g. 60%) "rainy"
 (decreased probability, e.g. 25%) "cloudy" (greatly
 decreased probability, e.g. 15%)

- At High Temperature (1.2):
 "sunny" (slightly reduced probability, e.g., 34%) "rainy"
 (slightly increased probability, e.g., 33%) "cloudy" (increased
 probability, e.g., 33%)

Ultimate example sentences:

- At low Temperature: ***"Weather is sunny today."*** (most likely)
- At high Temperature: ***"The weather is cloudy today."*** (more variation possible)

Similarities and Differences

Agreements:
- All three affect variation in output.
- Higher values usually lead to more diverse responses.
- Lower values give more predictable results.

Differences:
- Top K chooses a fixed number of options regardless of the probability distribution.
- Top P dynamically adapts to the specific probability distribution.
- Temperature works globally on all probabilities.

Top K, Top P and Temperature in AI text generation

Pros and Cons

Top K

Advantages:
- Easy to understand and implement.
- Prevents highly unlikely choices effectively.

Disadvantages:
- Less flexible under different probability distributions.
- May be too restrictive if K is set too low.

Top P

Advantages:
- Automatically adapts to different contexts.
- Provides a good balance between coherence and variety.

Disadvantages:
- Less intuitive to understand than Top K.
- May be too restrictive at very low values.

Temperature

Advantages:
- Intuitive to understand and set up.
- Gives fine control over overall "creativity.

Disadvantages:
- Can lead to inconsistency or excessive repetition at extreme values.
- Less adaptive to different contexts.

Practical Application

Choosing Settings

For factual information:
- Low Top K (10-50) or low Top P (0.1-0.3) or low Temperature (0.3-0.5)

For creative writing:
- Higher Top K (100-1000) or higher Top P (0.9-1.0) or higher Temperature (0.7-1.0)

For conversations:
- Moderate settings (Top K: 50-100, Top P: 0.7-0.9, Temperature: 0.6-0.8)

Combining Methods

You can use all three methods at the same time:
1. Top K first filters for a fixed number of options.
2. Top P then filters within that selection.
3. Finally, Temperature adjusts the probabilities within the remaining set.

This is like a librarian who first makes a fixed selection (Top K), then looks at the most relevant books within that selection (Top P), and finally decides how adventurous the final choice is (Temperature).

Experimentation and Finetuning

To find the best settings:
1. Start with default values (e.g., Top K 50, Top P 0.9, Temperature 0.7).
2. Generate multiple responses with different settings.
3. Evaluate results based on:
   - Relevance
   - Creativity
   - Coherence
   - Suitability for your specific task
4. Adjust gradually and repeat the process.

Top K, Top P and Temperature in AI text generation

Summary

Top K, Top P, and Temperature are powerful tools to control the output of AI text generation. By understanding and experimenting with them, you can tailor the AI assistant to your specific needs, whether generating factual reports, creative stories, or natural conversations.

Our ongoing example shows how these three methods work together:
1. Top K limits the choice to a fixed number of options (in our case, 5).
2. Top P further refines this by retaining only the most likely options (in our case, 3).
3. Temperature ultimately determines how "bold" the final choice is.

By combining these methods, you can gain very precise control over the generation process and tailor the output to your specific needs.

September 2024,

Anthony Loeff - @Litic.ai
https://github.com/anthonyloeff