Chapter NLP:IV

IV. Text Models

- Language Modeling
- Large Language Models
- Text Generation

Text Generation

Autoregressive language models generate text by iteratively predicting the next token.

1. Start with an initial sequence $\mathbf{x}_{<i}$. E.g. the start of sentence token <s>
2. At each step n: decode w_n according to the language model $P(\cdot \mid \mathbf{x}_{<n})$.
3. Append w_n to $\mathbf{x}_{<n}$ and go to 2 for step $n + 1$.

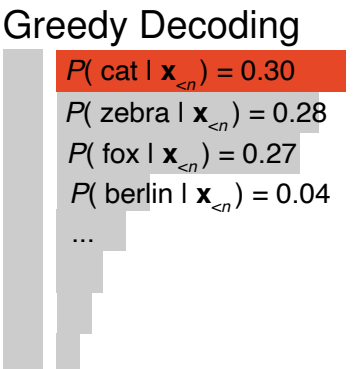
Text Generation

Decoding

The decoding strategy describes how x_n is chosen from $P(\cdot \mid \mathbf{x}_{<n})$.

- The simplest decoding strategy is greedy decoding: Always use the most probable token (argmax) at each step n.

$$x_n = \underset{i \in V}{\text{argmax}}\ p(x_n^i \mid \mathbf{x}_{<n})$$



Example (Mistral-7B; Ancestral Sampling; $\tau = 1.3$, $k = 40$):

The professor is a man of many talents. He is a professor of mathematics, a professor of physics, a professor of chemistry, a professor of biology, a professor of geology, a professor of astronomy, a professor of history, a professor of philosophy, a professor of literature, a professor of art, a professor of music, a professor of psychology, a professor of sociology, a professor of anthropology, a professor of economics, a professor of political science, a professor of law, a professor of religion, a professor of ethics, a professor of logic, a professor of rhetoric, a professor of linguistics, . . .

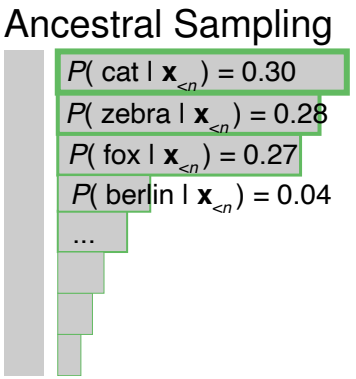
Text Generation

Decoding

The decoding strategy describes how x_n is chosen from $P(\cdot \mid \mathbf{x}_{<n})$.

- Better: sample x_n according to the distribution (ancestral sampling).

$$x_n^i \sim P(\cdot \mid \mathbf{x}_{<n})$$



Example:

The professor is currently working with a number of individuals within the University for the 2012 London Olympics, to help maximise athlete performance at the games.

This includes work with British Fencing, where she has a team of three fencers, led by former Olympian and Commonwealth Champion James-Andrew Davis with his team of sabre fencers in readiness for selection for the games.

Professor Hoggarth is also working closely with the British Taekwondo team, where one of her athletes, Bianca Walkden, is on the podium list for a medal in the event.

Text Generation

Decoding

Problems of text generated by language models:

1. Boring and repetitive text due to smoothing / maximum likelihood training.
2. Long tail of low-probability tokens can still be sampled.
3. Hard to recover from local minima.
4. Entropy of the task or typical documents:
   - Some tasks have few likely generations.
     `Translate to German: I love my cat.` `Ich liebe meine Katze.`
   - Frequent documents (of a genre) are often undesirable generations.
     `Generate a website about cats:` `404 Not Found.`
   - Most tasks have many "good" generations. When is a generation (not) correct?

Mitigate these issues by adapting the decoding:

- **Scoring:** Modify $P(\cdot \mid \mathbf{w}_{<i})$ to match the generation goal.
- **Strategy:** Given the scores, select tokens to find a good generation.

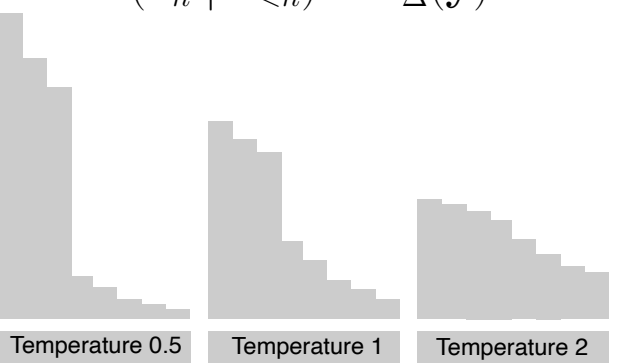
Text Generation

Scoring: Temperature

Scale the model output (probability mass) by a factor τ. Before the softmax.

- A $\tau < 1$ moves probability to frequent tokens
  $\rightarrow$ low surprisal text.

$$P(x_n \mid \mathbf{x}_{<n}) = \sigma_\Delta(\mathbf{y})^{\frac{1}{\tau}}$$



Example (Mistral-7B; Ancestral Sampling; $\tau = 0.7$, $k = 40$):

The professor is known for his work on the physics of soft matter.

Venkat G. Murthy, a professor in the College of Science and a materials science and engineering faculty member of the Charles E. Via Jr. Department of Civil and Environmental Engineering, has been named a fellow of the American Physical Society (APS).

Murthy was recognized for his work in the physics of soft matter, which has wide-ranging applications in materials engineering and the study of life. These applications include the self-assembly of block copolymers and the folding of proteins, among others.

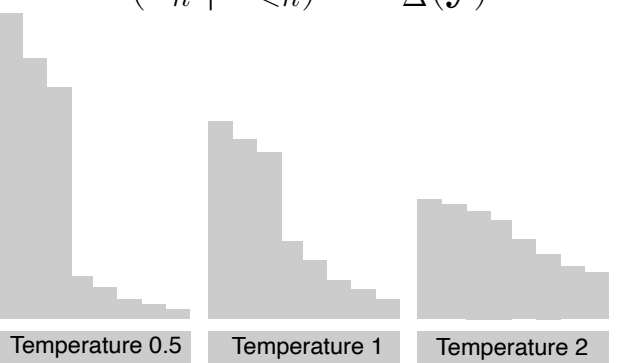
Text Generation

Scoring: Temperature

Scale the model output (probability mass) by a factor τ. Before the softmax.

- A $\tau < 1$ moves probability to frequent tokens
  $\rightarrow$ low surprisal text.
- A $\tau > 1$ moves probability to rare tokens
  $\rightarrow$ high surprisal text.

$$P(x_n \mid \mathbf{x}_{<n}) = \sigma_\Delta(\mathbf{y})^{\frac{1}{\tau}}$$



Example (Mistral-7B; Ancestral Sampling; $\tau = 1.3$; $k = 40$):

The professor is on the offensive with a vengeance! In this, the latest round in their titanic struggle, the great men of science are on your side, ready to help combat the wily Warlord and his sinister alien allies when they launch another fiendish bid for universal domination.

If they can only succeed... But as always with the warlike Warlord and his ally Omnister, time is NOT of the essence! Each game lasts ten thousand years. And a lot can go down between then and now.

To help the scientist in their struggle against the forces of Evil the Professor has now released the first volume of his seminal work . . .

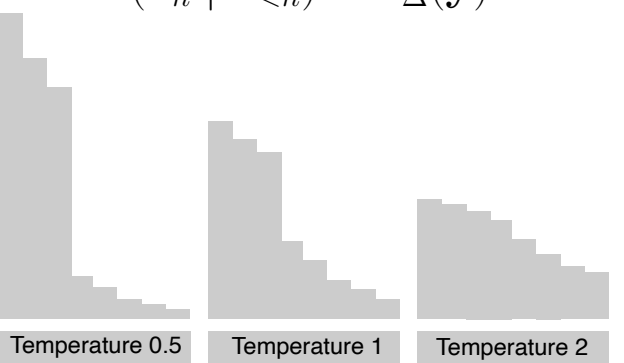
Text Generation

Scoring: Temperature

Scale the model output (probability mass) by a factor τ. Before the softmax.

- A $\tau < 1$ moves probability to frequent tokens $\rightarrow$ low surprisal text.
- A $\tau > 1$ moves probability to rare tokens $\rightarrow$ high surprisal text.
- Peaky distributions lead to repetitive text.

$$P(x_n \mid \mathbf{x}_{<n}) = \sigma_\Delta(\mathbf{y})^{\frac{1}{\tau}}$$



Example (Mistral-7B; Ancestral Sampling; $\tau = 0.3$; $k = 40$):

The professor is a man of many talents. He is a professor of English at the University of California, Berkeley, and he is also a writer. He has written many books, including The Professor and the Madman, which is about the life of the professor.

The professor is a man of many talents. He is a professor of English at the University of California, Berkeley, and he is also a writer. He has written many books, including . . .

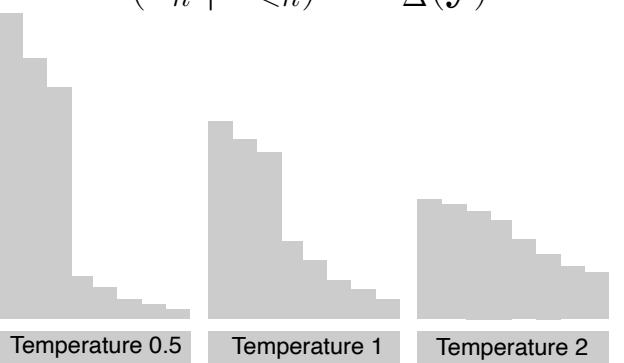
Text Generation

Scoring: Temperature

Scale the model output (probability mass) by a factor τ. Before the softmax.

- A $\tau < 1$ moves probability to frequent tokens $\rightarrow$ low surprisal text.
- A $\tau > 1$ moves probability to rare tokens $\rightarrow$ high surprisal text.
- Peaky distributions lead to repetitive text.
- Flat distributions lead to incoherent text.

$$P(x_n \mid \mathbf{x}_{<n}) = \sigma_\Delta(\mathbf{y})^{\frac{1}{\tau}}$$



Example (Mistral-7B; Ancestral Sampling; $\tau = 3$; $k = 40$):

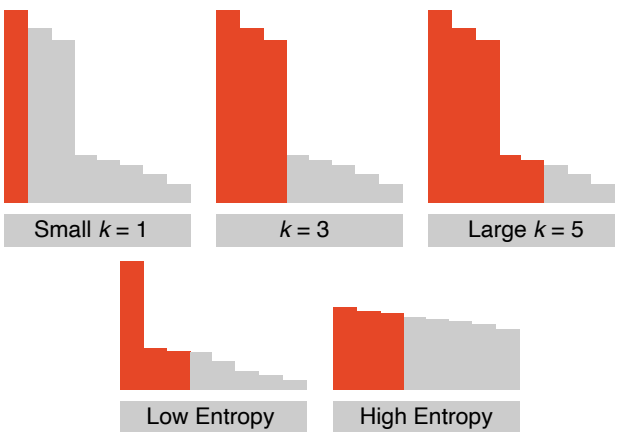
The professor is speaking a foreign dialect into her cell-phone handset. At best what sounds back like Russian must really fall within one half a planet west? and one hour in Moscow before our own Central day can find me. Who cares? ...

Text Generation

Scoring: Top-k Sampling

Truncate the long tail of low-probability tokens to prevent unlikely continuations. Keep only the k tokens with the highest probability.

- Large $k \rightarrow$ atypical, incoherent text.
- Small $k \rightarrow$ bland, repetitive text.
- But: For low-entropy distributions, we still sample tokens from the tail.
- But: For high-entropy distributions, we truncate high-probability tokens.



Example (Mistral-7B; Ancestral Sampling;$\tau = 1.1$; $k = 32000$):

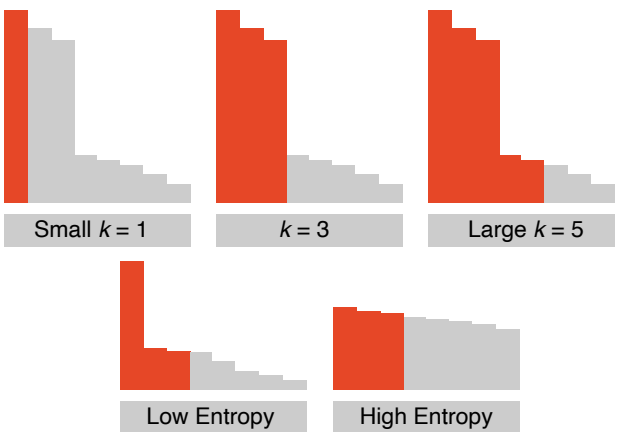
Apple CEO Tim Cook recently announced that everyone at Apple was "deeply sorry" for making and selling the iPod shuffle, an adorably useless knockoff of icomm's 1999 SeeBlue Mood, with its can-not speakers, built-in earbud-lookalikes, and circuitry, all jammed into a shell normally reserved for triangle plastic makeup cases. Wired reviews takes a look back ~~(rem: he's boring! He's GONE! Brian sucks!)~~at the SeeBlue Mood, a precursor ...

Text Generation

Scoring: Top-k Sampling

Truncate the long tail of low-probability tokens to prevent unlikely continuations. Keep only the k tokens with the highest probability.

- Large $k \rightarrow$ atypical, incoherent text.
- Small $k \rightarrow$ bland, repetitive text.
- But: For low-entropy distributions, we still sample tokens from the tail.
- But: For high-entropy distributions, we truncate high-probability tokens.



Example (Mistral-7B; Ancestral Sampling;$\tau = 1.1$; $k = 3$):

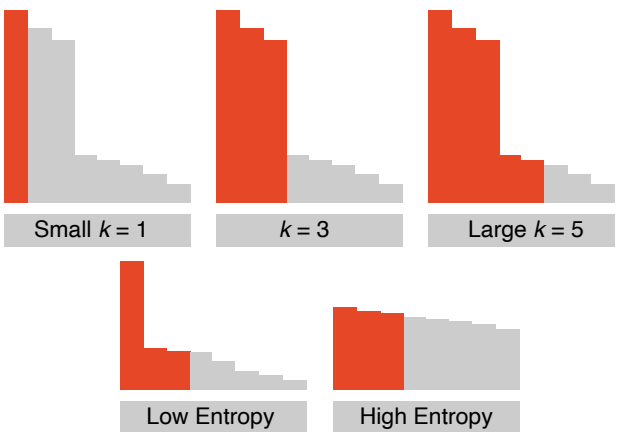
Apple CEO Tim Cook recently announced a $200 million expansion of the company's Austin facility, and now the tech giant has confirmed that it will also expand in San Diego. The expansion is part of a larger effort by Apple to invest $30 billion in the US over the next five years. The San Diego expansion will include a 250,000 sq. ft. facility, according to the San Diego Union-Tribune.

Text Generation

Scoring: Top-k Sampling

Truncate the long tail of low-probability tokens to prevent unlikely continuations. Keep only the k tokens with the highest probability.

- Large $k \rightarrow$ atypical, incoherent text.
- Small $k \rightarrow$ bland, repetitive text.
- But: For low-entropy distributions, we still sample tokens from the tail.
- But: For high-entropy distributions, we truncate high-probability tokens.



Example (Mistral-7B; Ancestral Sampling;$\tau = 1.1$; $k = 40$):

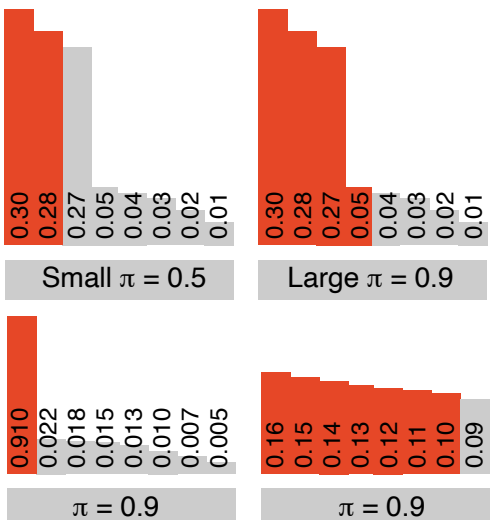
Apple CEO Tim Cook recently announced his company's next iPod Touch and iPod Nano players, while also detailing some pretty major changes to its software. The most important change is that you can now download all of the 3,000 native Apple applications directly to the iPhones and iPods without any syncing.

Text Generation

Scoring: Nucleus (Top-p) Sampling

Truncate the long tail of low-probability tokens to prevent unlikely continuations. Keep the most probable tokens up to a cumulative probability mass π.

- Small $\pi \rightarrow$ bland, repetitive text (as with k).
- Fails for peaky distributions with one very frequent and many valid but less frequent tokens.



Example (Mistral-7B; Ancestral Sampling;$\tau = 1.1$; $k = 32000$; $\pi = 0.9$):

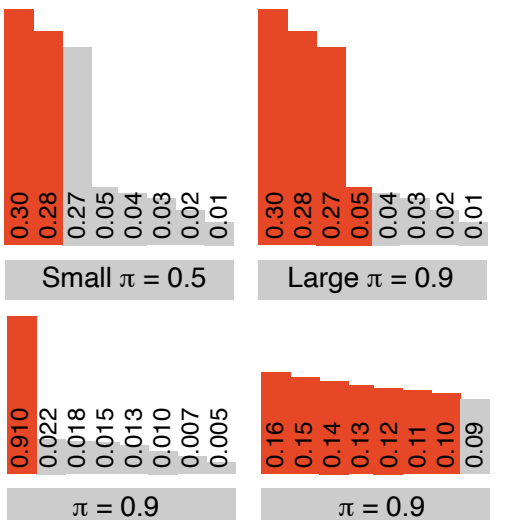
Apple CEO Tim Cook recently announced that the company would start manufacturing MacBook Pros in the US, which is great news. But is Apple doing it for the right reasons? Not quite. This comes from Steve Jobs biographer Walter Isaacson, who argued that the tech company's "crusade" for more and more manufacturing jobs in the US is not about "benevolence" but because of business interests.

Text Generation

Scoring: Nucleus (Top-p) Sampling

Truncate the long tail of low-probability tokens to prevent unlikely continuations. Keep the most probable tokens up to a cumulative probability mass π.

- Small $\pi \rightarrow$ bland, repetitive text (as with k).
- Fails for peaky distributions with one very frequent and many valid but less frequent tokens.



Example (Mistral-7B; Ancestral Sampling;$\tau = 1.1$; $k = 32000$; $\pi = 0.9$):

The company's CEO Donald Trump Jr. said at a news conference ...
The president Donald Trump's next budget, expected to be ...
The secretary Donald trump of energy has handed a ...
Former sneaker salesperson Donald Trump Jr. took to Twitter Saturday to remind constituents ...

Text Generation

Strategy: Contrastive Search [Su, 2022] [greedy decoding]

Generated text often degenerates quickly (dull; repetitive tokens or phrases). Contrastive decoding punishes repetitive tokens with a penalty.

$$x_n = \underset{i \in V}{\text{argmax}} \; (1-\alpha) \cdot P(x_n^i \mid \mathbf{x}_{<n}) - \alpha \cdot \max(d_{\text{Cosine}}(i, x_j) : \forall_{j<n})$$

- The degeneration penalty is the highest cosine similarity between a token in the distribution (x_n^i) and any previously generated token in $\mathbf{x}_{<n}$.
- The α balances the model confidence and the penalty for the argmax.
- High α with greedy decoding leads to incoherent text.

Example (Mistral-7B; Greedy Search; $k = 10$; $\alpha = 0$):

Apple CEO Tim Cook recently announced that the company will be releasing a new version of the iPhone in September. The new iPhone will be called the iPhone 5S and will be available in three colors: black, white, and gold. The iPhone 5S will have a new A7 processor, which is twice as fast as the A6 processor in the iPhone 5. It will also have a new M7 motion coprocessor, which will allow the phone to track your movements and activity. The iPhone 5S will also have a new camera, which will be able to take better pictures in low light. It will also have a new fingerprint sensor, which will allow you to unlock your phone with your fingerprint.

Text Generation

Strategy: Contrastive Search [Su, 2022] [greedy decoding]

Generated text often degenerates quickly (dull; repetitive tokens or phrases). Contrastive decoding punishes repetitive tokens with a penalty.

$$x_n = \underset{i \in V}{\text{argmax}} \; (1-\alpha) \cdot P(x_n^i \mid \mathbf{x}_{<n}) - \alpha \cdot \max(d_{\text{Cosine}}(i, x_j) : \forall_{j<n})$$

- The degeneration penalty is the highest cosine similarity between a token in the distribution (x_n^i) and any previously generated token in $\mathbf{x}_{<n}$.
- The α balances the model confidence and the penalty for the argmax.
- High α with greedy decoding leads to incoherent text.

Example (Mistral-7B; Greedy Search; $k = 10$; $\alpha = 0.5$):

Apple CEO Tim Cook recently announced that the company will donate $1 million to the American Red Cross to help victims of Hurricane Harvey. Cook made the announcement on Twitter, saying the money will go to the Greater Houston Community Fund at the Houston Community Foundation.

Apple joins a slew of companies that have pledged money to help the victims of flooding in southeast Texas. AT&T, ExxonMobil, Southwest Airlines, H-E-B supermarkets, Valero Energy Corp.'s VLO, -0.08% Ben & Jerry's ice cream . . .

Text Generation

Strategy: Contrastive Search [Su, 2022] [greedy decoding]

Generated text often degenerates quickly (dull; repetitive tokens or phrases). Contrastive decoding punishes repetitive tokens with a penalty.

$$x_n = \underset{i \in V}{\text{argmax}} \; (1-\alpha) \cdot P(x_n^i \mid \mathbf{x}_{<n}) - \alpha \cdot \max(d_{\text{Cosine}}(i, x_j) : \forall_{j<n})$$

- The degeneration penalty is the highest cosine similarity between a token in the distribution (x_n^i) and any previously generated token in $\mathbf{x}_{<n}$.
- The α balances the model confidence and the penalty for the argmax.
- High α with greedy decoding leads to incoherent text.

Example (Mistral-7B; Greedy Search; $k = 10$; $\alpha = 0.8$):

Apple CEO Tim Cook recently announced at WWPD (Worldwiden developers conferences) in San Fransiscothat they?rereleasaing 10.9 MaverseR.I will tell yu a lil abou this release so stay tuned Apps and iPaDevelopeR are inforcED to ue IOS STYLIStics in thi raleA sEleCtIon fRO MUltiPIE oTioN WIL bE pROVIdeD tO uS eVErYtIme wE aDd i . . .

Remarks:

- It is a common observation that generated text degenerates the longer it gets. The suspected reason for this is *forced exposure*(or *teacher forcing*): During training, the models is only exposed to real data. Although the training is also autoregressive (appending word-by-word), the model output is always discarded and replaced by a word from the (human-written) training data. During predicting, the model's input is it's output from the previous iteration. [He, 2021]

Text Generation

Strategy: Beam Search [iterative decoding]

Iteratively sampling only the next token often leads to low probability sequences or local minima. Beam search decodes k hypotheses (*beams*) in parallel.

1. Start with an initial sequence $\mathbf{x}_i$. Decode k continuations $x_i^1, \ldots, x_i^k$.
2. At each step n: decode k continuations $x_n^1, \ldots, x_n^k$ for each of the k continuations of the previous step $n-1$.
3. Determine the sequence probability $P(\mathbf{x}_n^k)$ for each of the $2 \cdot k$ beams.
4. Prune the k lowest scoring beams. Go to 2.

Example (Mistral-7B; Greedy Decoding; $k = 40$; $\alpha = 0.2$; 5 Beams):

Apple CEO Tim Cook recently announced the company's plans to invest $350 billion in the U.S. economy over the next five years.

"We have a deep sense of responsibility to give back to our country and the people who help make our success possible," Cook said. "We're focusing our investments in areas where we can have a direct impact on job creation and job preparedness. We have a long way to go, and we're going to continue accelerating our pace of giving."

Remarks:

- There are many other sampling and decoding strategies.
  - Microstat: Sample to approximate the perplexity of human text. [basu, 2020]
  - Contrastive Decoding: Avoid weak predictions by discounting the probabilities of an expert language model by the probabilities of an amateur model [Li, 2023]
  - η-sampling: Truncate relatively and absolutely improbable words [Hewitt, 2022]

- In practice, a mixture of various techniques are used:
  - k is often set high (10-100) to eliminate the (nonsensical) long tail.
  - Beam search and ancestral sampling are often combined.
  - A low degeneration penalty (α 0.1–0.3) can often improve results, especially for more creative text.
  - The temperature is used with ancestral sampling to control creativity and surprisal of the text.

- Large models are complex systems. How various parameters act on individual prompts is often elusive.