Generative LLMs

CS 485, Fall 2024
Applications of Natural Language Processing

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

Generative & **L**arge **LM**s

- "Generative": can sample/choose textual output from the model
- "Large": billions of parameters, trillions of tokens of training data (web, books, etc.)
  - "LLM" is ambiguous term

- Model: typically left-to-right Transformer

- Two pretty different variants!
  - 1. Pure language modeling
  - 2. + Instruction tuning: encourage useful responses

- Some aspects of them today
  - Decoding/sampling
  - Instruction tuning, fine tuning
  - Open-weight API: https://api.together.xyz/playground/
    - Or use Huggingface

- Left-to-right LM training (next word prediction) as "massively multi-task" learning

Prefix {choice_1, choice_2}	**Task**
In my free time, I like to {run, banana}	Grammar
I went to the zoo to see giraffes, lions, and {zebras, spoon}	Lexical semantics
The capital of Denmark is {Copenhagen, London}	World knowledge
I was laughing the entire time, the movie was {good, bad}	Sentiment analysis
The word for “pretty” in Spanish is {bonita, hola}	Translation
First grade arithmetic exam: 3 + 8 + 4 = {15, 11}	Math question

Prefix	**Next word [task]**
A transformer is a deep learning architecture, initially proposed in	2017 [factual recall]
A transformer is a deep learning architecture, initially proposed in 2017	, [comma prediction]
A transformer is a deep learning architecture, initially proposed in 2017,	that [grammar]
A transformer is a deep learning architecture, initially proposed in 2017, that	relies [impossible task?]

[from Jason Wei, Nov 2023]

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

[GPT-2 paper (2018)]

Training data

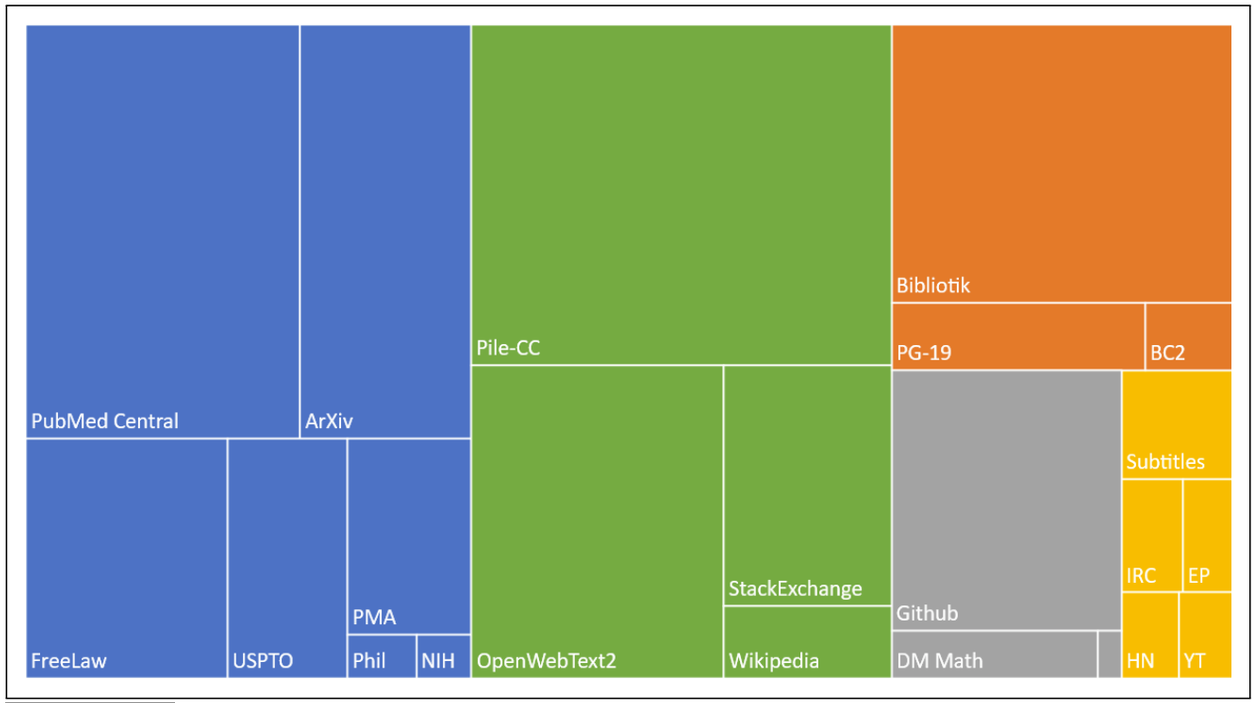

Figure 10.5 The Pile corpus, showing the size of different components, color coded as academic (articles from PubMed and ArXiv, patents from the USPTA; internet (webtext including a subset of the common crawl as well as Wikipedia), prose (a large corpus of books), dialogue (including movie subtitles and chat data), and misc . Figure from Gao et al. (2020).

- see also WaPo / AI2 analysis https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/
- Data collection is very uncontrolled; quality filters etc. have huge impact. Many unknowns still in the training data

$p(w_{N+1} \mid w_1..w_N) =$

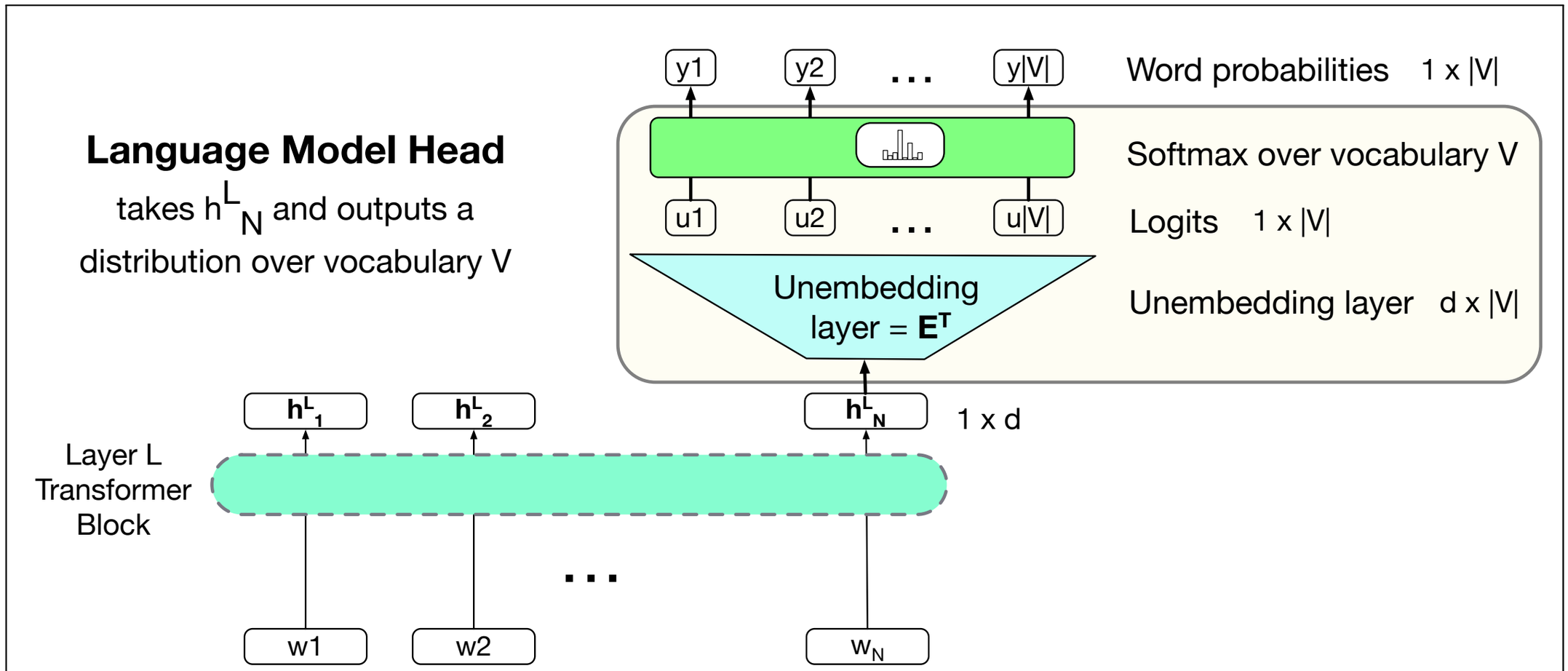

Figure 10.13 The language modeling head: the circuit at the top of a transformer that maps from the output embedding for token N from the last transformer layer ($\mathbf{h}^L_N$) to a probability distribution over words in the vocabulary V.

- Left-to-right, a.k.a. causal model
- attend only to left a.k.a. "masked attention" a.k.a."decoder-only"

[*SLP3 reading*]

Context window

- Self-attention isn't practical to run on an entire document
  - Why?
- Context window limits
  - GPT-2 (2018): 512-1024 tokens
    - similarly BERT, RoBERTA (2018-2019)
  - Llama 3 (2024): 8192 tokens
  - rising fast...
- Current research: Are long-range contexts used *well*? How to improve?

- How to select a full output string, from an iterative next-word-prob model?
  - Greedy decoding
  - Random sampling
  - Beam search (heuristic optimization)
- But....

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

Figure 1: Even with substantial human context and the powerful GPT-2 Large language model, Beam Search (size 32) leads to degenerate repetition (highlighted in blue) while pure sampling leads to incoherent gibberish (highlighted in red). When $b \geq 64$, both GPT-2 Large and XL (774M and 1542M parameters, respectively) prefer to stop generating immediately after the given context.

- Most popular solutions:
  - Top-K and Top-P (*nucleus sampling*) filtering
  - *Temperature* to control variability

Holtzman et al., 2020

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Sampling, *t*=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

Nucleus, *p*=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Figure 3: Example generations continuing an initial sentence. Maximization and top-k truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

Holtzman et al., 2020

Multitask learning, explicit datasets:
Just train the LM for *all* tasks
p(desired answer | input text)

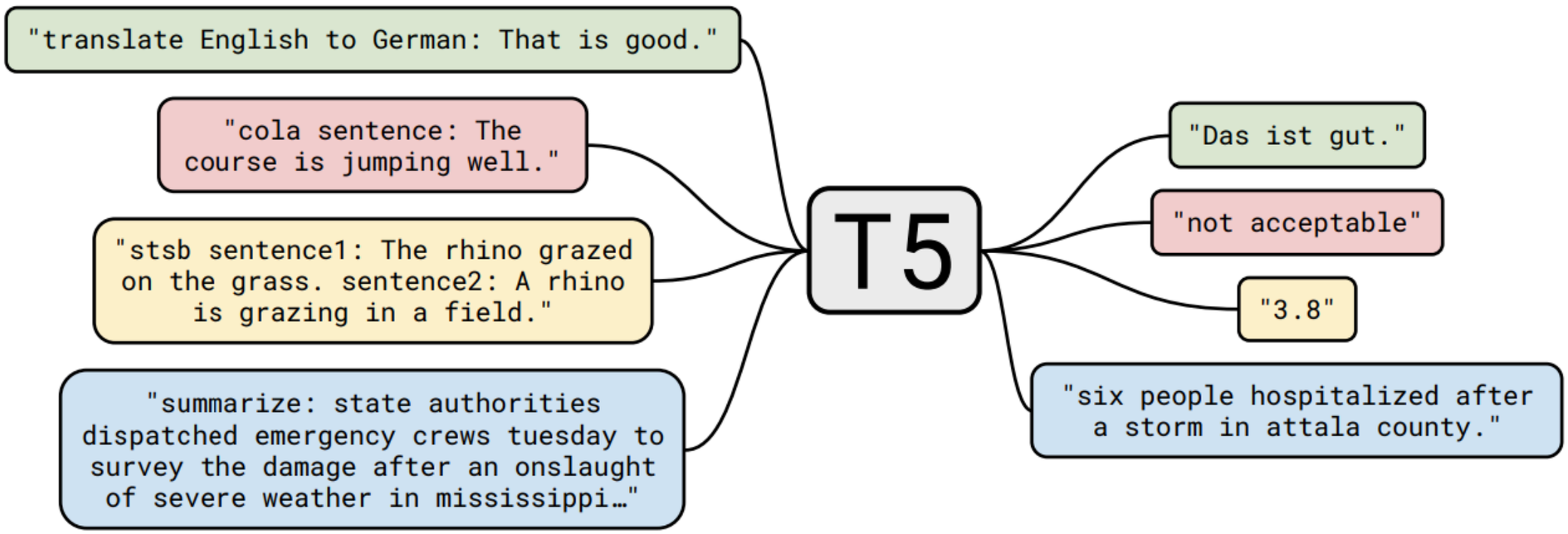

Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

- T5 is a "encoder-decoder" model
- Fine-tuned T5 is popular for supervised tasks

[Raffel et al. 2020]

LoRA fine-tuning

- BERT fine-tuning was for all ~100M params
- But what about larger gen LLMs? 7B, 70B, etc?
- **Lo**w-**R**ank **A**daptation

$$\mathbf{h} = \mathbf{x}\mathbf{W} + \mathbf{x}\mathbf{A}\mathbf{B} \tag{10.14}$$

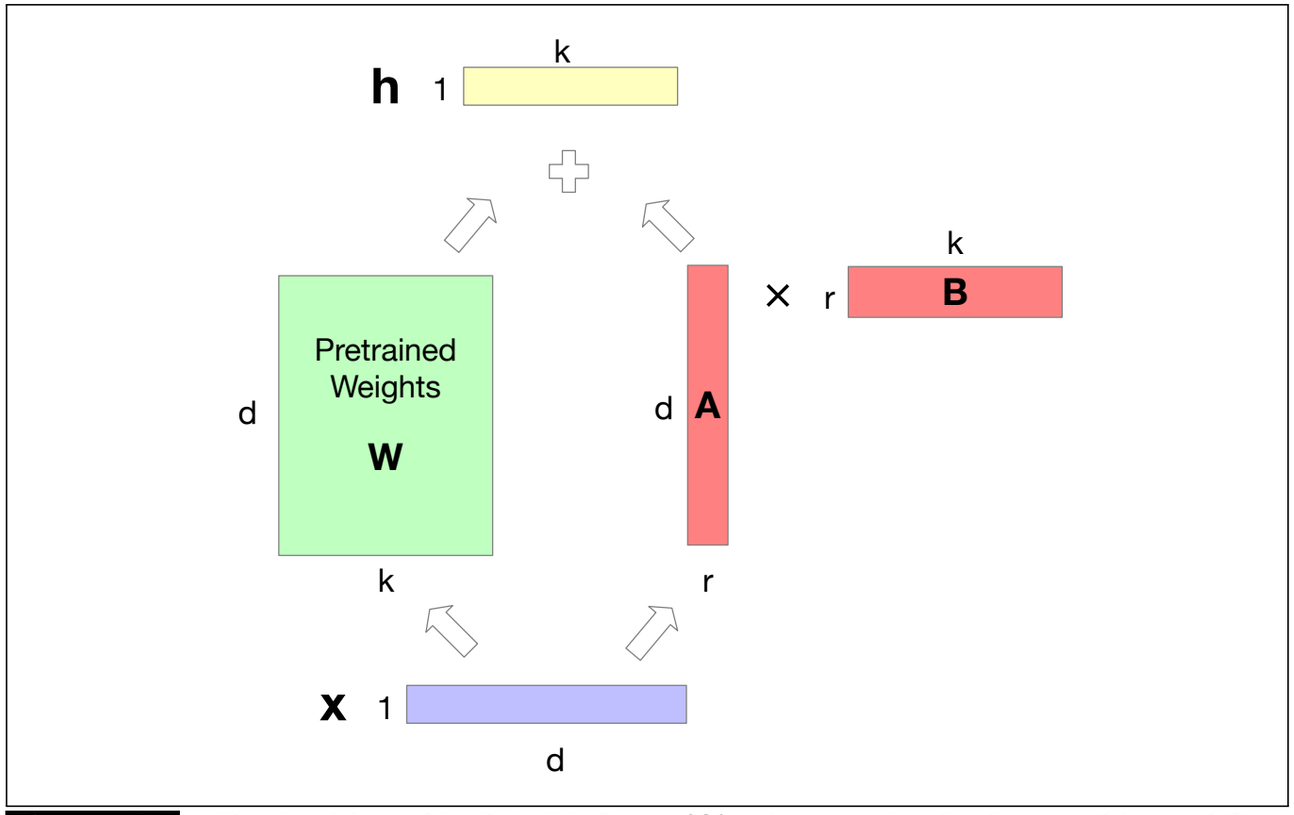

Figure 10.8 The intuition of LoRA. We freeze **W** to its pretrained values, and instead fine-tune by training a pair of matrices **A** and **B**, updating those instead of **W**, and just sum **W** and the updated **AB**.