Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Com funciona el ChatGPT i altres IA conversacionals

Pol Garcia Recasens

CROMAI group - BSC & UPC
pol.garcia@bsc.es

12/11/2024

Facultat d'Informàtica de Barcelona

Agenda

1. Què és un chatbot
2. Aplicacions pràctiques d’un chatbot amb IA
3. IA conversacional vs IA generativa
4. Large Language Models
5. Entrenament d'un model generatiu
6. Optimitzant la inferència
7. Batching
8. Altres tècniques

Què és un chatbot

Un **chatbot** es un programa informàtic que simula una conversa humana amb un usuari final.

Chatbots tradicionals

- Respostes preprogramades, arbres de decisió.

Chatbots amb IA

- Utilitzen tècniques d'aprenentatge automàtic, NLP i NLU.
- Interpreten amb precisió les preguntes dels usuaris i les associen a intencions específiques.

Agents virtuals avançats

- Poden automatitzar tasques i generar contingut (text, imatges, etc).
- Combinen aquestes tecnologies d'IA amb l'automatització robòtica de processos (RPA).

Què és un chatbot

Un **chatbot** es un programa informàtic que simula una conversa humana amb un usuari final.

Chatbots tradicionals

“Quin temps farà demà?” El chatbot respon dient que plourà.

Chatbots amb IA

“Com pinta el temps de demà?” El chatbot, interpretant la pregunta, respon dient que plourà.

Agents virtuals avançats

“Com pinta el temps de demà?” —l’agent virtual no només interpreta i prediu la pluja de demà, sinó que ofereix configurar una alarma per avisar possibles retards als trens.

Evolució dels chatbots a agents virtuals

Els **assistents virtuals** intel·ligents poden:

Entendre converses lliures amb models de llenguatge avançats.

Automatitzar tasques rellevants.

Basats en la IA conversacional i l'auto-aprenentatge però també l'automatització robòtica de processos.

Exemples: Siri d'Apple, Alexa d'Amazon, Gemini de Google, ChatGPT d'OpenAI.

El valor dels chatbots

Assistència instantània. Responen ràpidament a preguntes dels usuaris sense necessitat d'intervenció humana.

Ús generalitzat. Presents en altaveus intel·ligents, aplicacions de missatgeria i entorns laborals com Slack.

Capacitats avançades d'IA. Els assistents virtuals intel·ligents poden entendre converses i automatitzar tasques.

Integració. Els xatbots es poden integrar en eines com Microsoft Teams per fomentar la col·laboració.

Orquestració de fluxos de treball. Permet fluxos de treball complexos en temps real (e.g gestió de tasques de CRM).

Estratègies de màrqueting. Chatbots d'IA disponibles 24/7 per a una experiència personalitzada i consistent a través de diferents canals digitals.

Avantatges

Millora del compromís del client i fidelitat de marca

Gestionen interaccions 24/7, eliminant esperes i optimitzant el servei

Augmenten la satisfacció i la fidelitat dels clients.

Reducció de costos i augment de l'eficiència operativa

Estalvi significatiu comparat amb la contractació de personal dia i nit.

Responen les preguntes repetitives, permetent que els agents humans es dediquin a casos complexos i ajudant a escalar el suport segons la demanda.

Generació de leads i satisfacció del client

Poden qualificar leads i donar suport en vendes, responent preguntes sobre productes i serveis en el moment.

Per compres més complexes, poden connectar el client directament amb un agent de vendes.

Casos d’ús

Interactuen amb aplicacions mòbils, termòstats intel·ligents i electrodomèstics connectats.

A les empreses, els chatbots s’utilitzen en màrqueting, atenció al client, equips d’IT i HR, i centres de contacte.

Capacitats Avançades

Memòria conversacional: els chatbots recorden i incorporen el context de les interaccions anteriors.
Automatització (RPA): permet realitzar tasques complexes i transferir fàcilment la conversa a un agent humà amb el context complet.

Usats en apps de missatgeria social, plataformes propietàries, webs, aplicacions, i fins i tot per telèfon (IVR).

Exemples

Assistència 24/7 en atenció al client o recursos humans.
Recomanacions personalitzades en e-commerce.
Màrqueting de productes i serveis.
Gestió de cites en salut i recordatoris automatitzats.

IA conversacional vs IA generativa

La IA generativa pot generar respostes automàtiques a partir de la base de coneixement de l'organització.

Chatbots amb IA Conversacional

Interpreten preguntes/comentaris i generen respostes semblants a les humanes.

https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ceo-generative-ai

IA conversacional vs IA generativa

La IA generativa pot generar respostes automàtiques a partir de la base de coneixement de l'organització.

Chatbots amb IA Conversacional

Interpreten preguntes/comentaris i generen respostes semblants a les humanes.

Chatbots amb IA Generativa

Van un pas més enllà generant nou contingut, com text, imatges i so.

Reconeixement, resum, traducció, predicció i creació de contingut en resposta a consultes d'usuari.

Tot això sense necessitat d'intervenció humana, gràcies als models de llenguatge de gran escala (**LLM**).

Tendència: El 85% dels directius prediuen la interacció directa de la IA generativa amb clients en els pròxims dos anys ("The CEO's guide to generative AI" de l'IBV).

https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ceo-generative-ai

Large Language Models

Què són els LLMs? Models d'IA entrenats amb grans quantitats de dades, capaços d'entendre i generar llenguatge natural i altres tipus de contingut per resoldre múltiples tasques.

Els LLMs han estat crucials per portar la IA generativa al centre d'atenció pública i impulsar la seva adopció empresarial en diverses funcions i casos d'ús.

Exemples destacats: OpenAI (GPT-3, GPT-4), Meta (Llama), Google (BERT, PaLM) i IBM (Granite).

Capacitats: inferir context, generar respostes coherents, traduir, resumir, respondre preguntes, generació de codi.

Els LLMs són models "fundacionals," que ofereixen flexibilitat per a diferents aplicacions sense necessitat de models específics per cada cas.
Amb milers de milions de paràmetres capturen patrons complexos.

Pre-entrenant LLMs

Aprenentatge auto-supervisat. Un model de llenguatge entrenat per a la modelització causal del llenguatge pren una seqüència de tokens de text com a entrada i retorna la distribució de probabilitat per al següent token.

Diferents families de models

- Generative Pre-trained Transformer (**GPT**)
- Open Pre-trained Transformer (**OPT**)
- Large Language Model Meta AI (**LLaMA**)
- …

Entrenats amb un gran corpus de dades (increasing!)

- GPT (2018): ≈ 5 GB (BookCorpus)
- OPT (2022): ≈ 2000 GB (BookCorpus + Stories + CCNews + ThePile + Pushshift.io Reddit)

Pre-entrenant LLMs

Mètode d’entrenament que no necessita dades etiquetades manualment.

ChatGPT aprèn a identificar patrons lingüístics i a predir el següent fragment o paraula dins d'un text donat.

Recopilació de dades (textos de diverses fonts com llibres, articles, webs).

ChatGPT intenta predir paraules o frases amagades en frases, com per exemple: "El cel és ___".

A través d'aquest procés, aprèn estructures gramaticals i patrons de significat.

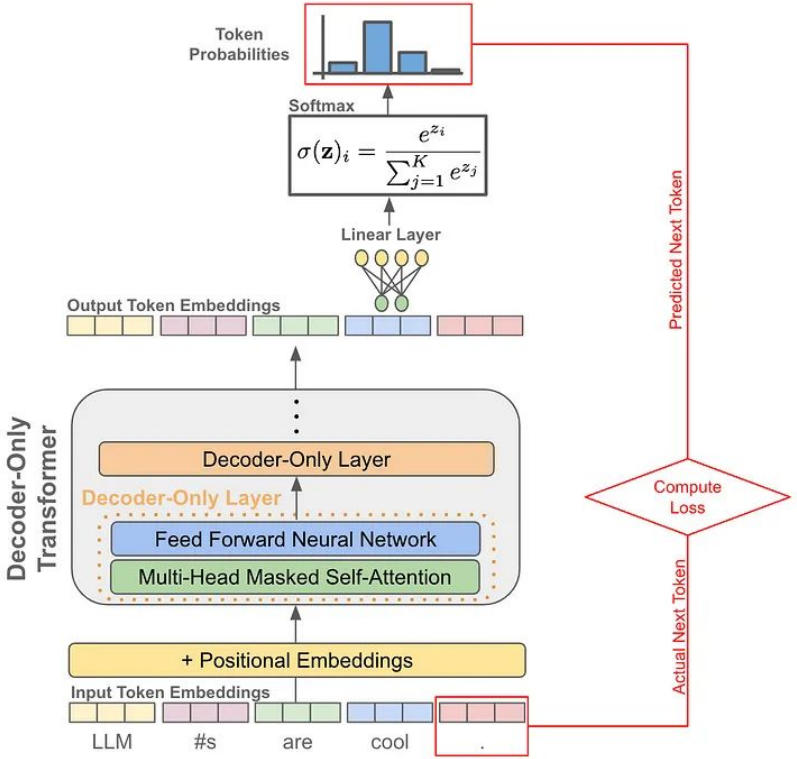
Si s’equivoca, s'ajusta a través d'un procés anomenat backpropagation per millorar en futures prediccions.

Escalabilitat: permet entrenar models amb grans volums de dades sense etiquetes.

Generalització: aprèn matisos de llenguatge que es poden aplicar a múltiples tasques.

Eficiència: redueix la necessitat de dades etiquetades manualment, que són costoses i lentes de crear.

Pre-entrenant LLMs



https://medium.com/@akash.kesrwani99/understanding-next-token-prediction-concept-to-code-1st-part-7054dabda347

Pre-entrenant LLMs

GPT (2018): 117M *o* 234 MB.

GPT-2 (2019): 1.5B *o* 3 GB.

GPT-3 (2020): 175B *o* 350 GB.

OPT-175B (mateix tamany que GPT-3) s'ha entrenat amb 992 80GB A100 GPUs.

GPT-4 (2023): Els rumors diuen que GPT-4 té 1,76 bilions de paràmetres.

L'arquitectura real de GPT-4 és un mixture model amb 220 mil milions de paràmetres dividits en 8 conjunts de paràmetres.

Arquitectura de Mixture of Experts (MoE).

https://pub.towardsai.net/gpt-4-8-models-in-one-the-secret-is-out-e3d16fd1eee0

Self-attention

L'**autoatenció** permet entendre el context d'una paraula dins d'una frase.

Aquesta tècnica ajuda el model a veure quines paraules de la frase són importants per a cada paraula.

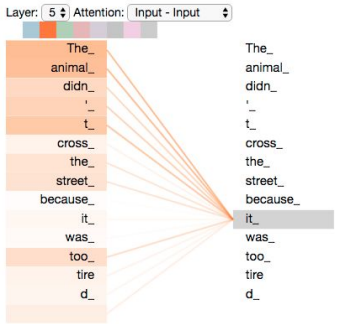
Si tenim la frase "El gat menja peix," el model pot associar la paraula "gat" amb "menja" i "peix".

Incorpora informació anterior quan processa la paraula actual.

Però hi ha un detall important…

Quan el Transformer genera text, cada posició de la frase actual pot "veure" totes les posicions anteriors i la seva pròpia posició.

Els estats interns s'han de mantenir entre iteracions (**KV cache**) per evitar re-computacions.



Shazeer, Noam. "Fast transformer decoding: One write-head is all you need." *arXiv preprint arXiv:1911.02150* (2019).

Com posar un chatbot amb IA generativa en producció

1. Entrenament i optimització del model

   Entrena i ajusta el model d'IA generativa (p. ex., ChatGPT) amb les dades específiques.

   Assegura't que respongui de manera adequada i compleixi amb els requeriments de l'empresa.

Com posar un chatbot amb IA generativa en producció

1. Entrenament i optimització del model

    Entrena i ajusta el model d'IA generativa (p. ex., ChatGPT) amb les dades específiques.

    Assegura't que respongui de manera adequada i compleixi amb els requeriments de l'empresa.

2. Desplegament en un servidor

    Carrega el model en un servidor que escali (al núvol o en un servidor dedicat).

    Crea una API que permeti als clients accedir al chatbot des de diferents aplicacions (web, mòbil, etc.).

    Aquesta API actua com a "pont" entre els clients i el servidor que executa el model d'IA.

Com posar un chatbot amb IA generativa en producció

1. Entrenament i optimització del model

   Entrena i ajusta el model d'IA generativa (p. ex., ChatGPT) amb les dades específiques.

   Assegura't que respongui de manera adequada i compleixi amb els requeriments de l'empresa.

2. Desplegament en un servidor

   Carrega el model en un servidor que escali (al núvol o en un servidor dedicat).

   Crea una API que permeti als clients accedir al chatbot des de diferents aplicacions (web, mòbil, etc.).

   Aquesta API actua com a "pont" entre els clients i el servidor que executa el model d'IA.

3. Clients fent peticions

   Els clients interactuen amb el chatbot a través d'aplicacions com una web o aplicacions de missatgeria.

   Cada vegada que un client envia un missatge, aquest passa per l'API fins al servidor on està allotjat el model.

   El servidor processa cada petició, utilitzant l'IA generativa per analitzar la sol·licitud i generar una resposta.

   La resposta es retorna a l'aplicació client a través de l'API per ser mostrada a l'usuari en temps real.

Posant en producció chatbots amb IA generativa

Els usuaris fan submit de una petició a un servidor d'inferència (API connectada amb un servidor amb el LLM).

Ens interessa maximitzar el **throughput** (req/s) del sistema i minimitzar la **latència** de l'usuari (s).

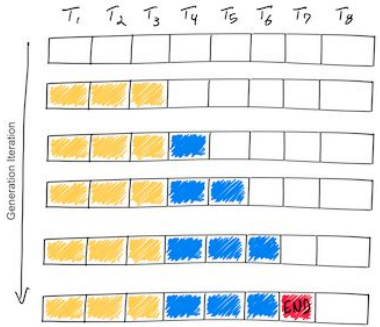
Latència: TTFT (temps que triga a generar el primer token després de processar la petició), TPT (temps mitjà que triga a generar cada token).

Throughput: Número d'output tokens que un servidor d'inferencia pot generar per tots els usuaris i peticions.

Fases de la generació

Els LLMs prediuen un sol token a la vegada.

Generació autoregressiva: Procediment d'inferència que consisteix a cridar iterativament el model amb els seus propis resultats generats, a partir d'uns inputs inicials.



Per a cada petició:

Es comença amb una seqüència de tokens (anomenada prefix o prompt).

"Quina és la capital de Califòrnia: "

El LLM produeix una seqüència de tokens, aturant-se només després de produir un token d'aturada o d'assolir una longitud màxima de seqüència.

["S", "a", "c", "r", "a", "m", "e", "n", "t", "o"].

https://www.anyscale.com/blog/continuous-batching-llm-inference

KV cache

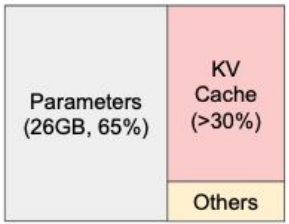
The KV cache of a token can be calculated as:

2 (FP16) * 2 (key and value tensors) * (hidden dimensions) * (number of layers)

In the case of OPT 1.3B, a request with 512 input tokens and 256 output tokens requires 50 MB.

A batch of 512 requests demands at most **25.6 GB** of memory space to store the KV cache.

Therefore, the KV cache size grows with the model size and the number of tokens per request.



Fases de la generació

Fase de prompt

A partir d'un prompt (x1,...,xn), es generen les keys (k1, …, kn) i values (v1,…,vn).

Els tokens del prompt són coneguts, es pot paral·lelitzar el càlcul per a cada token.

Aquesta fase es pot computar de manera eficient, cada posició es processa en paral·lel.

Fases de la generació

Fase de prompt

A partir d'un prompt (x1,...,xn), es generen les keys (k1, …, kn) i values (v1,…,vn).

Els tokens del prompt són coneguts, es pot paral·lelitzar el càlcul per a cada token.

Aquesta fase es pot computar de manera eficient, cada posició es processa en paral·lel.

Fase de generació autoregressiva

Genera un token per iteració fins a acabar la generació o arribar a un límit establert.

La computació no es pot paral·lelitzar a causa de les dependències entre tokens.

Requereix productes matriu-vector per a cada iteració.

Aquesta fase tendeix a infrautilitzar els recursos de la GPU, degut a la naturalesa seqüencial del procés de generació.

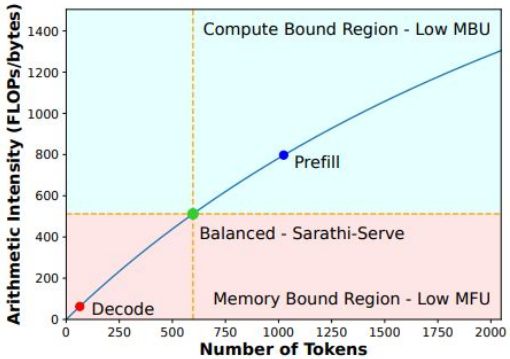
Factors limitants en la generació

El rendiment de un pas d'inferència (generació d'un token) pot ser:

Limitat per I/O de memòria: el rendiment està limitat pel temps d'accés a la memòria.

Limitat per la computació: el rendiment està limitat pel temps dedicat a les operacions de càlcul.

La mètrica utilitzada per mesurar el factor limitant és la intensitat aritmètica, definida com la relació entre les operacions de càlcul i els bytes transferits des de la memòria HBM.



Agrawal, Amey, et al. "Taming throughput-latency tradeoff in llm inference with sarathi-serve." *arXiv preprint arXiv:2403.02310* (2024).

Factors limitants en la generació

Baixa intensitat aritmètica en la generació autoregressiva

- El temps dedicat a carregar els paràmetres del model des de la memòria és superior al temps dedicat a calcular les operacions.
- Això provoca una infrautilització dels recursos de la GPU.
- Operacions matriu-vector en inferència de batch únic (pocs FLOPs).

Una manera d'augmentar la intensitat aritmètica és calcular múltiples peticions amb una mateixa càrrega de paràmetres del model (**batching**).

Augmentar la intensitat aritmètica millora el rendiment del sistema.

Tècniques de batching

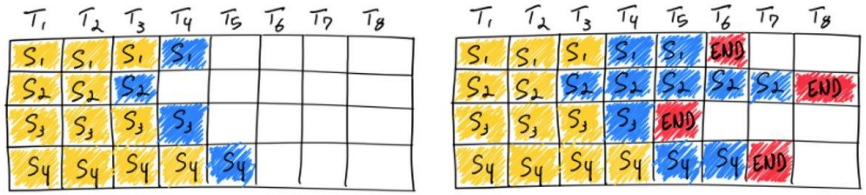
Dos nivells de granularitat

- Request-level granularity
- Iteration-level granularity

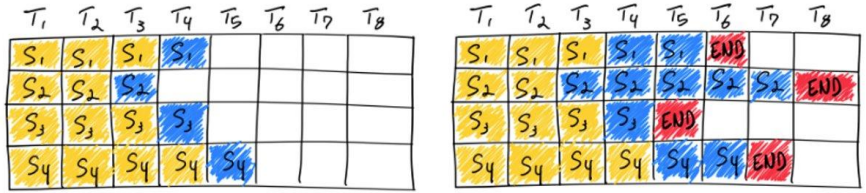
Tres tipus de batching

- Static batching
- Dynamic batching
- Continuous batching

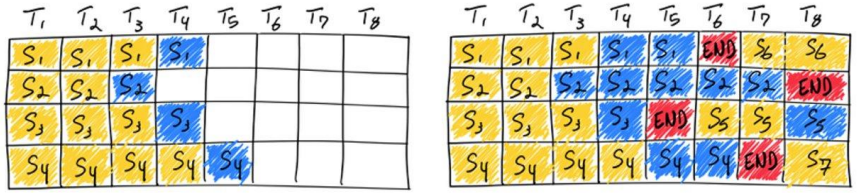
https://www.anyscale.com/blog/continuous-batching-llm-inference



Static batching



Dynamic batching



Continuous batching

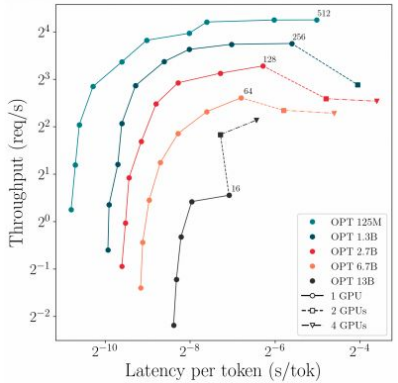
Factors limitants en el batching

Incrementar la mida batch size augmenta la càrrega computacional.

Si el temps de càlcul és més gran que el temps de memòria-IO (temps de càrrega dels pesos), arribem a un límit de rendiment.

Després d'un cert punt, augmentar la mida del lot no millora més el rendiment.

PERÒ... Com podem agrupar diverses peticions durant la inferència? No sabem el nombre de tokens de sortida, com gestionem el cache KV?



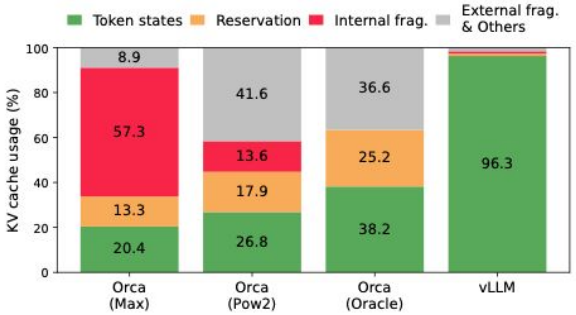
Recasens, Pol G., et al. "Towards Pareto Optimal Throughput in Small Language Model Serving." *Proceedings of the 4th Workshop on Machine Learning and Systems*. 2024.

PagedAttention

L'algoritme **PagedAttention** identifica i resol de manera eficient les fragmentacions de memòria en la gestió de la memòria cache KV.

Inspirat en conceptes tradicionals de sistemes operatius, com el paging i la memòria virtual.

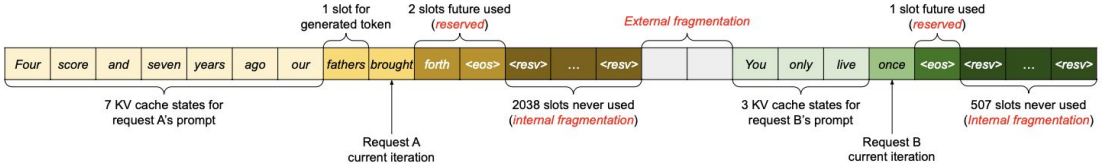
Permet que la memòria cache KV sigui no contigua, mitjançant l'assignació de memòria en pàgines o blocs de mida fixa.



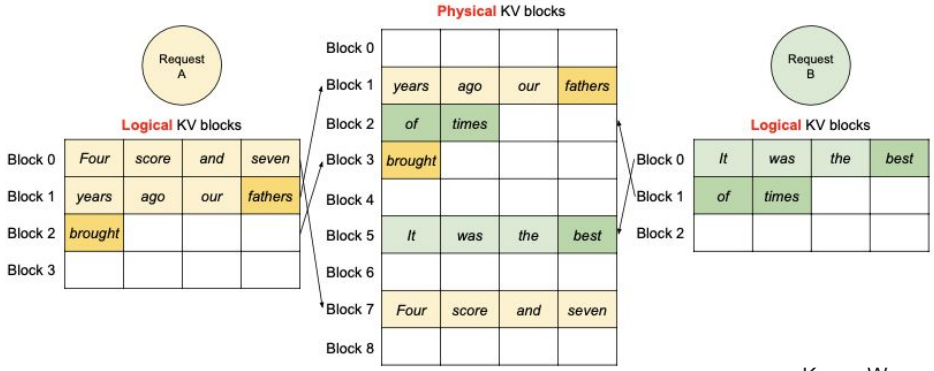
Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." *Proceedings of the 29th Symposium on Operating Systems Principles*. 2023.

PagedAttention

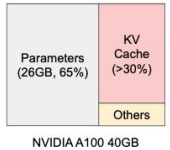
PagedAttention permet guardar blocs de memòria en espai de memòria no contigus.



Continuous batching



Continuous batching with Paged Attention



Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." *Proceedings of the 29th Symposium on Operating Systems Principles*. 2023.

Altres tècniques d’optimització

Quantització: minimitza l'ús de memòria comprimint els pesos del model.

Sparsity: elimina tokens i attention heads poc importants per reduir la complexitat computacional.

Offloading: tècniques que aprofiten la memòria de la CPU i el disc en escenaris de servei offline, per alleujar la càrrega de la GPU.

Sistemes d’inferència SOTA: Sarathi-serve, DeepSpeed-FastGen, vLLM, TGI, ORCA, AlpaServe, FlexGen.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Com funciona el ChatGPT i altres IA conversacionals

Pol Garcia Recasens

CROMAI group - BSC & UPC
pol.garcia@bsc.es

12/11/2024

Facultat d'Informàtica de Barcelona

Self-attention

A neural attention function takes a single **query**-vector *q* and a set of m different (**key**-vector, **value**-vector) pairs (represented by the matrices K and V), and produces an output vector z.

The output z is computed as a weighted sum of the different value vectors, where the weights are derived by comparing the query to the keys.