Estado del arte de Modelos de Lenguaje

Modelos

Un modelo de Machine Learning es una ecuación o fórmula matemática que tiene variables "vacías" o sin valor (llamados parámetros), las cuales se completan en la fase de entrenamiento, a través de un método de optimización.

Las redes neuronales son una arquitectura o "meta modelo", que permite crear diferentes modelos, a través del uso de los llamados "hiper parámetros" que crean diferentes versiones de una red neuronal.

Según el patrón de conexión de las neuronas, las redes neuronales se clasifican en diferentes "arquitecturas", como las redes neuronales convolucionales, recurrentes, y ahora último los transformers.

Tipos de modelos de lenguaje modernos

Existen varios tipos de modelos de lenguaje en el mundo open source. Las principales arquitecturas basadas en Transformers conforman los principales 3 tipos de modelos (creados el 2019, 2020 y 2021 respectivamente).

- BERT: Para tareas simples, como crear embeddings, MASK inference (adivinar la palabra en medio) y QA cortas
- MT5: Para tareas de traducción, summarization y QA corto (un poco más largo que BERT)
- LLM (Large Language Model) / SLM (Small Language Model): modelos basados en la arquitectura GPT
    - Mono-modelos: Un único modelo, la forma típica
    - Mixture Of Experts (MoE): Un modelo compuesto de varios modelos expertos en ciertos temas cada uno. Es más eficiente. Se puede "forzar" la expertiz de cada submodelo, o bien dejarlo a la red que escoja la expertiz.

Cuantización

Los modelos pueden estar en su "forma base" como fueron entrenados, la mayoría usa formato "HF" (coma flotante de 16 bits), o bien se pueden usar "cuantizados" para que pesen menos (los

lobotimiza un poco, es decir les quita rendimiento/accuracy), en números enteros de 8, 6, 5, 4, 3 o 2 bits. En general de 8 bits se considera que pierde muy poco rendimiento, casi nada, y hasta 5 bits es recomendable, 4 bits pierde mucha información y 2 bits habla muchas tonterías (sólo para hacer pruebas experimentales, no para uso en producción).

Ejemplos de modelos fundacionales LLM Open Source más famosos

A continuación modelos de lenguaje por empresas:

Modelo	Idioma	Empresa	RoPe	Parámetros	Peso HF*	Link
Falcon	Multi	Emiratos Árabes	2048	7B	15GB	https://huggingface.co/tiiuae/falcon-7b
Falcon-40B	Multi	Emiratos Árabes	2048	40B	84GB.	https://huggingface.co/tiiuae/falcon-40b
Llama2-7B	Inglés	Facebook	4k	7B	13.5GB	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
Llama2-13B	Inglés	Facebook	4k	13B	26GB	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
Llama2-70B	Inglés	Facebook	4k	70B	150GB	https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
Llama3-8B	Multi	Facebook	8k	8B		https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama3-70B	Multi	Facebook	8k	70B		https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Gemma-2b	Inglés	Google		2.51B		https://huggingface.co/google/gemma-2b-it
Gemma-7b	Inglés	Google		8.54B		https://huggingface.co/google/gemma-7b-it
Gemma2-9b						
Gemma2-27b						
Phi2	Inglés	Microsoft	2048	2.78B		https://huggingface.co/microsoft/phi-2
Phi-3-mini	Principalmente inglés	Microsoft	4k o 128k	3.82B		https://huggingface.co/microsoft/Phi-3-mini-4k-instruct o https://huggingface.co/microsoft/Phi-3-mini-128k-instruct
Phi-3-small	Principalmente inglés	Microsoft	8k o 128k	7.39B		https://huggingface.co/microsoft/Phi-3-small-8k-instruct o https://huggingface.co/microsoft/Phi-3-small-128k-instruct
Phi-3-medium	Principalmente inglés	Microsoft	4k o 128k	14B		https://huggingface.co/microsoft/Phi-3-medium-4k-instruct o https://huggingface.co/microsoft/Phi-3-medium-128k-instruct
Phi-3-vision	Principalmente inglés	Microsoft	128k	4.15B		https://huggingface.co/microsoft/Phi-3-vision-128k-instruct
Nemotron-340B	50 idioma	Nvidia	4096	340B		https://huggingface.co/nvidia/Nemotron-4-340B-Instruct o

	s y 40 lenguajes de código					https://huggingface.co/nvidia/Nemotron-4-340B-Reward
Mistral-0.2	Inglés	MistralAI		7B	15GB	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
Mistral-0.3	Multi	MistralAI		7B		https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
Mixtral	Multi5	MistralAI		8x7B	95GB (full 32 bits)	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
Mixtral	Multi	MistralAI		8x22B (141)		https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1
CommandR	Multi10	Coherence		35B	69.7 GB	https://huggingface.co/CohereForAI/c4ai-command-r-v01
DBRX		Databricks	32768	132B (16 expertos y elige 4)		https://huggingface.co/databricks/dbrx-instruct
Grok		Twitter	8,192	314B		https://huggingface.co/xai-org/grok-1
SmolLM-135M		HuggingFace		135M		https://huggingface.co/HuggingFaceTB/SmolLM-135M-Instruct
SmolLM-360M		HuggingFace		360M		https://huggingface.co/HuggingFaceTB/SmolLM-360M-Instruct
SmolLM-1.7B		HuggingFace		1.7B		https://huggingface.co/HuggingFaceTB/SmolLM-1.7B-Instruct
Mistral-Nemo	Multi	MistralAI		12B		
Codestral	Programming	MistralAI				
Mathstral		MistralAI				
DCLM		Apple				

Otros modelos, de uso académico o experimental:

Modelo	Idioma	Principal cualidad	RoPe	Parámetros	Peso HF*	Link
Tinyllama	Sólo inglés	Académico, es un modelo muy liviano. Trained on 90 days using 16 A100-40G GPUs on 3 trillion tokens		1.1B	2.2GB	https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
Tinyllama	Español	Experimental		1.1b	2.2GB	https://huggingface.co/bololab/tinyllama-spanish_16bit
MoE Tinyllama 3x1.1B	Multi26	Experimental		3x1.1B	5.2GB	https://huggingface.co/NickyNicky/Mix_TinyLlama-3x1B_oasst2_chatML_Cluster_3_2_1_V1

Mixtral Q8	Multi5	Versión cuantizada del original		8x7B	49.62 GB (52.12 GB for inference)	https://huggingface.co/TheBloke/Mixtral-8x7B-v0.1-GGUF
MiniCPM-Llama3-V 2.5	Inglés y Chino	Versión con visión, de una empresa China		8.54		https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5
Jamba	Multi?	No usa Transformers sino otra arquitectura llamada Mamba	256K	51.6B	102GB	https://huggingface.co/ai21labs/Jamba-v0.1
JetMoE	Multi	MIT, Princeton, y otras, Académico, Entrenado con datos de muy buena calidad, alcanza el rendimiento de llama2 gastando 0.08Millones en su entrenamiento (96 H100 por 2 semanas)	4096	8B (MoE con 2.2 active)	17GB	https://huggingface.co/jetmoe/jetmoe-8b-chat
Flor-1.3B	Ing-esp-cat	Basado en BLOOM para español y catalán		1.3B		https://huggingface.co/projecte-aina/FLOR-1.3B-Instructed
Flor-6.3B	Ing-esp-cat	Basado en BLOOM para español y catalán		6.3B		https://huggingface.co/projecte-aina/FLOR-6.3B-Instructed
Aguila-7B	Ing-esp-cat	Basado en Falcon para español y catalán		7B		https://huggingface.co/projecte-aina/aguila-7b

*Peso HF es en 16 bits float (cuantizado en Q8 (entero de 8 bits) pesa aproximadamente la mitad, en Q4 un cuarto, etcétera).

Ejemplos de modelos BERT y mt5 más útiles, seleccionados

Nombre	Autor	Función	Comentarios	Link
BETO	DCC Uchile	Existen diversos modelos para feature extraction (por palabra), Fill MASK, entre otros	Es un referente y usado como modelo fundacional para hacer FT	https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased o https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
Tulio	DCC Uchile	Fill MASK	Entrenado con español general y Chileno además	https://huggingface.co/dccuchile/tulio-chilean-spanish-bert
Patana	DCC Uchile	Fill MASK	Entrenado sólo con español	https://huggingface.co/dccuchile/patana-chilean-spanish-bert

			Chileno (noticias, webs, reclamos y tweets)	
NV-embed	Nvidia	Sentence Transformers	Se puede intencionar el embedding	https://huggingface.co/nvidia/NV-Embed-v1
Fast-Embed		Sentence Transformers	Librería	https://github.com/qdrant/fastembed
Sentence Similarity Spanish	HilamSid	Sentence Transformers	La he usado con buenos resultados, tiene miles de descargas mensuales	https://huggingface.co/hiiamsid/sentence_similarity_spanish_es
nomic-embed-text-v1.5	Nomic-AI	Sentence Transformers	Redimensionable	https://huggingface.co/nomic-ai/nomic-embed-text-v1.5
PubMed BERT Embeddings		Sentence Transformers	Hecho en base a pubmed	https://huggingface.co/NeuML/pubmedbert-base-embeddings
The crispy sentence embedding family from mixedbread ai.	**mixedbread-ai**	Sentence Transformers		https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1
GliNER	Estudiante phd en LIPN (París)	NER	Detecta Entidades dada una categoría	https://huggingface.co/urchade/gliner_multi-v2.1
Bert2Bert	Un emprendedor	Summarization	FT de BETO	https://huggingface.co/mrm8488/bert2bert_shared-spanish-finetuned-summarization

Finetunning (Supervised Fine tunning, SFT)

Finetunning es el proceso de tomar un modelo fundacional, como los de arriba, y entrenarlo para que siga determinado idioma, estilo, conocimiento.

- Full sin optimizaciones: No recomendado
- PEFT: Parameter Eficient Fine Tuning, entrenar sólo unos pocos parámetros, no todos.

    - LoRa: Descompone la matriz de pesos en 2 matrices más pequeñas. Entrena esas 2 matrices pequeñas y luego el resultado del modelo es la suma de estas 2 matrices multiplicadas + el modelo original
    - QLoRa: LoRa pero cuantizando la matriz de pesos.
    - Para más detalles recomiendo leer
        - https://aman.ai/primers/ai/parameter-efficient-fine-tuning/#
        - https://huggingface.co/docs/peft/main/en/conceptual_guides/lora
- MEFT: Memory Eficient. Técnicas para disminuir la memoria
    - Optimizer cuantización: Cuantizar los parámetros del optimizador, que sin modificar pesan aprox lo mismo que el modelo original.
    - Gradient ranking: usar LoRa sobre la matriz de gradientes, permite entrenar Full-parameter usando apenas un 25% más de memoria que el modelo, es muy prometedor el método, se llama **GaLore** (paper**: https://arxiv.org/pdf/2403.03507)**
    - LOMO: Un método que reduce el 10% de memoria, no he leído cómo

Para información MUY técnica visitar:
https://huggingface.co/docs/transformers/perf_train_gpu_one

Librerías para hacer Fine Tunning

- Llama Factory (Parece que es la más fácil, es con interfaz web y un Docker)
- Unsolth (tiene versión gratis y de pago)
- Axolotl (pasó de moda parece)
- Ludwig
- AutoTrain (La oficial de Hugging Face)

Casos de ejemplo

- Usando LOMO: Full parameter 65B model on a single machine with 8 RTX 3090, each with 24GB memory.
- Mixtral 8x7B en LoRa con modelo Q4:
    - It only requires **28GB** to fine-tune the **8x7B** model with LLaMA Factory.
    - https://huggingface.co/mistralai/Mixtral-8x7B-v0.1/discussions/10
- Mixtral 8x7B LoRa Q4
    - A100 40GB
    - https://www.reddit.com/r/LocalLLaMA/comments/18gwoke/fine_tuning_mixtral_8x7b/
- Llama Factory (una librería para hacer SFT), se puede optimizar aún más con técnicas MEFT, pero aquí una muestra de requerimiento de memoria estimados (tomar como el máximo, se puede mejorar)

Hardware Requirement

** estimated*

Method	Bits	7B	13B	30B	70B	8x7B
Full	AMP	120GB	240GB	600GB	1200GB	900GB
Full	16	**60GB**	120GB	300GB	600GB	400GB
GaLore	16	16GB	32GB	64GB	160GB	120GB
Freeze	16	20GB	**40GB**	80GB	200GB	160GB

LoRA	16	16GB	32GB	**64GB**	160GB	120GB
QLoRA	8	10GB	20GB	40GB	80GB	**60GB**
QLoRA	4	6GB	12GB	24GB	48GB	30GB
QLoRA	2	4GB	8GB	16GB	24GB	18GB

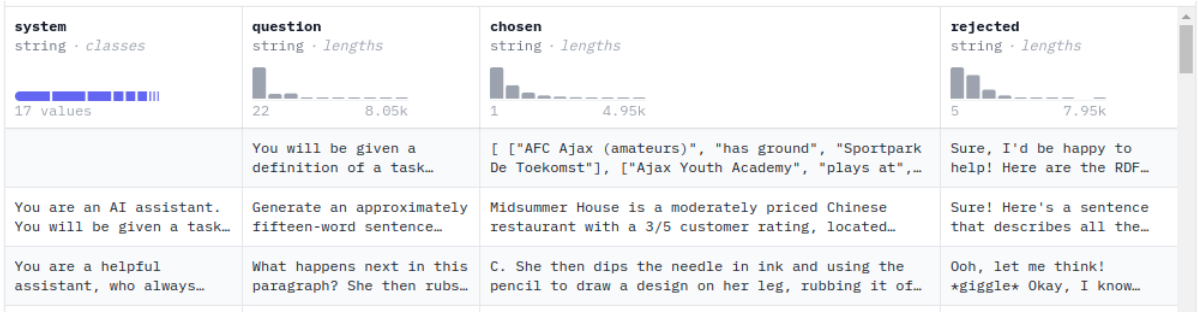
- Unsloth (librería para hacer LoRa muy prometedora)
    - **0% loss in accuracy** - no approximation methods - all exact.
    - Supports 4bit and 16bit QLoRA / LoRA finetuning via bitsandbytes.
    - DPO SUpport
    - Mistral7B en sólo 12.4GB de VRAM!!!
    - CodeLlama 34B con sólo 27.4GB VRAM!!!
    - https://github.com/unslothai/unsloth
- Axolotl (otra librería para hacer SFT)
    - https://github.com/OpenAccess-AI-Collective/axolotl

Optimización de preferencias

La optimización de preferencias permite que el modelo escoja la mejor respuesta dado un set de pregunta-respuesta.

Para información técnica: https://huggingface.co/blog/pref-tuning (recomendable leer hasta “Links”, después es MUY técnico todo)

- DPO (Direct preference optimization): Requiere un dataset con una pregunta – una respuesta elegida, y una o más respuestas rechazadas.



system string · *classes* 17 values	question string · *lengths* 22 8.05k	chosen string · *lengths* 1 4.95k	rejected string · *lengths* 5 7.95k
	You will be given a definition of a task…	[["AFC Ajax (amateurs)", "has ground", "Sportpark De Toekomst"], ["Ajax Youth Academy", "plays at",…	Sure, I'd be happy to help! Here are the RDF…
You are an AI assistant. You will be given a task…	Generate an approximately fifteen-word sentence…	Midsummer House is a moderately priced Chinese restaurant with a 3/5 customer rating, located…	Sure! Here's a sentence that describes all the…
You are a helpful assistant, who always…	What happens next in this paragraph? She then rubs…	C. She then dips the needle in ink and using the pencil to draw a design on her leg, rubbing it of…	Ooh, let me think! *giggle* Okay, I know…

- IPO (Identity Preference Optimisation): Método optimizado de DPO
- KTO (Kahneman-Tversky Optimisation): permite usar ejemplos etiquetados con manito arriba y manito abajo (for example, the 👍 or 👎 icons one sees in chat UIs) por lo que es más fácil aplicarlo en la práctica.