Análise de Eficiência de InfiniAttention na Predição de Tokens em Português

Fernando Gubitoso Marques
Matheus Rodrigues de Souza Félix

7 de novembro de 2024

Resumo

Neste estudo exploratório, implementamos e utilizamos a arquitetura InfiniAttention para o modelo GPT-2 e treinamos o modelo para predição de tokens sobre um dataset de textos enciclopédicos em português. Comparamos seu desempenho com o modelo GPT-2 com *self-attention* tradicional. Por fim, realizamos algumas alterações ao modelo com InfiniAttention em relação ao utilizado no comparativo para tentar melhorar seu desempenho. Fornecemos um exemplo de geração textual deste último modelo a partir da predição de tokens.

1 Introdução

O tamanho do contexto para um modelo de linguagem é um fator que limita o desempenho de modelos que lidam com grandes cadeias de tokens. Para manter a coesão entre o texto pré-existente e o texto a ser gerado, os grandes modelos de linguagem empregam janelas contextuais de milhares de tokens. No entanto, isto pode incutar custos computacionais e *overheads* de memória elevados. Abordando esta questão está a arquitetura Infini Attention.[2] Esta tecnologia permite o armazenamento de informações a respeito de tokens que já sairam da janela de contexto, efetivamente extendendo a "memória"do modelo e permitindo a ele atentar a informações passadas.

Neste estudo, será apresentado um módulo de InfiniAttention e uma comparação do seu desempenho em uma tarefa de predição textual em português brasileiro em relação ao módulo tradicional de auto-atenção *multihead* no modelo de linguagem GPT-2.

2 Implementação

Seguindo a implementação original do Infini Attention[2], o Multi-Head Attention (MHA) tem a atenção computada $\mathbf{A}_{\text{dot}} \in \mathbb{R}^{N \times d_{\text{value}}}$ a partir de uma sequência de segmentos $\mathbf{X} \in \mathbb{R}^{N \times d_{\text{model}}}$ a partir das seguintes etapas.

Primeiro, são computadas as projeções da entrada em query, key e value:

$$\mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \quad \mathbf{Q} = \mathbf{X}\mathbf{W}_Q. \tag{1}$$

$\mathbf{W}_K \in \mathbb{R}^{d_{\text{model}} \times d_{\text{key}}}$, $\mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{value}}}$ e $\mathbf{W}_Q \in \mathbb{R}^{d_{\text{model}} \times d_{\text{key}}}$ são matrizes de projeção com parâmetros aprendidos durante o treinamento. Em sequência, a atenção é computada como:

$$\mathbf{A}_{\text{dot}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{\text{model}}}}\right)\mathbf{V} \tag{2}$$

No Infini Attention, os valores de query, key e value ($\mathbf{Q}$, $\mathbf{K}$ e $\mathbf{V}$) são reutilizados entre o processamento dos segmentos de entrada através de uma matriz associativa com o objetivo de possibilitar atualizações e recuperações de valores como uma projeção linear e melhorar a estabilidade do treino.

No Infini Attention, recuperamos o novo conteúdo $\mathbf{A}_{\text{mem}} \in \mathbb{R}^{N \times d_{\text{value}}}$ da memória $\mathbf{M}^{s-1} \in \mathbb{R}^{d_{\text{key}} \times d_{\text{value}}}$ usando a query $\mathbf{Q} \in \mathbb{R}^{N \times d_{\text{key}}}$ como:

$$\mathbf{A}_{\text{mem}} = \frac{\sigma(\mathbf{Q})\mathbf{M}_{s-1}}{\sigma(\mathbf{Q})\mathbf{z}_{s-1} + \varepsilon} \tag{3}$$

Neste trecho, σ e $\mathbf{z}_{s-1} \in \mathbb{R}^{d_{\text{key}}}$ são, respectivamente, uma função de ativação não linear e um termo de normalização. Seguindo a implementação original, são registradas a soma de todas as

keys como o termo de normalização $\mathbf{z}_{s-1}$ e usamos a função de ativação ELU + 1 element-wise. Em contraste com a implementação original, foi adicionado um termo ε, com valor padrão 10^{-6}, no denominador para fins de estabilidade numérica.

$$\mathbf{M}_s \leftarrow \mathbf{M}_{s-1} + \sigma(\mathbf{K})^T\mathbf{V}, \quad \mathbf{z}_s \leftarrow \mathbf{z}_{s-1} + \sum_{t=1}^{N} \sigma(\mathbf{K}_t) \tag{4}$$

Os estados de memória $\mathbf{M}_s$ e $\mathbf{z}_s$ são passados para o processamento do segmento $s+1$, criando uma recorrência em cada camada do cálculo de atenção. A implementação original não conta com definições sobre a inicialização dos estados de memória e termos de normalização. Além disso, não há registro relacionado ao aprendizado ou não desses valores durante a etapa de treinamento. Neste estudo foi implementada a inicialização com valores zerados para $\mathbf{M}$ e valores aleatórios multiplicados pelo fator 10^{-10} para os fatores de normalização $\mathbf{z}$. Ambos são reinicializados a cada etapa de inferência para evitar o acúmulo exagerado de gradientes e a passagem de informação que exceda o texto de entrada durante o treinamento.

Para agregar a atenção local $\mathbf{A}_{\text{dot}}$ com a atenção de memória $\mathbf{A}_{\text{mem}}$, é utilizado um escalar β aprendido durante o treinamento.

$$\mathbf{A} = \text{sigmoid}(\beta) \odot \mathbf{A}_{\text{mem}} + (1 - \text{sigmoid}(\beta)) \odot \mathbf{A}_{\text{dot}} \tag{5}$$

Com o uso da biblioteca `transformers` e PyTorch, foi criada uma classe `GPT2InfiniAttention` a partir da classe base `GPT2Attention`. Instanciando um novo modelo `GPT2LMHeadModel`, que conta com uma camada de classificação no topo do modelo para a tarefa de previsão de próximo token, todos os módulos `GPT2Attention` foram substituídos pelo `GPT2InfiniAttention`. Para contornar algumas problemáticas relacionadas a interfaces fora de contexto do trabalho, este presente estudo também implementou as abordagens propostas em um módulo `InfiniTransformer` [2] implementado exclusivamente em PyTorch.

3 Metodologia

3.1 *Dataset*

O dataset empregado neste estudo é uma fração do "Corpus Geral do Português Brasileiro Contemporâneo"(Carolina) [1], desenvolvido e mantido pela Universidade de São Paulo. Este corpus abrange um vasto volume de textos em português brasileiro contemporâneo, de 1970 a 2021, detalhando a proveniência e tipologia dos textos. Especificamente, utilizamos a seção "wik"do corpus, que é composta predominantemente por textos enciclopédicos.

A versão utilizada do corpus, Ada 1.2, disponível desde 8 de março de 2023, inclui aproximadamente 823 milhões de tokens. Dentre estes, a seção "wik"contribui com 403.927.145 tokens, representando a maior parcela do corpus. Para este estudo, selecionou-se apenas 50% e 10% da seção "wik", onde os textos foram concatenados, tokenizados e divididos em chunks igual ao tamanho do contexto do modelo. Após isso, foi realizada uma divisão de 98% para treinamento e 2% para validação.

3.2 Métricas

Para a avaliação do desempenho dos modelos neste estudo, utilizamos inicialmente a entropia cruzada (CrossEntropy), também utilizada como função de custo. A entropia cruzada é uma medida comum em tarefas de modelagem de linguagem que quantifica a diferença entre duas distribuições de probabilidade. A função é definida para classificação de múltiplas classes como:

$$\text{CrossEntropy}(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$$

onde y é o vetor de etiquetas verdadeiras e $\hat{y}$ é o vetor das probabilidades preditas pelo modelo.

Complementarmente, a perplexidade (PPL) foi utilizada para acompanhar o desempenho do modelo. A perplexidade é derivada da entropia cruzada e é calculada pela fórmula:

$$\text{PPL}(y, \hat{y}) = \exp(\text{CrossEntropy}(y, \hat{y}))$$

A perplexidade aumenta à medida que as previsões do modelo divergem das etiquetas verdadeiras, fornecendo uma medida direta da qualidade das previsões do modelo em termos de incerteza.

3.3 Tarefa

O objetivo principal deste estudo é avaliar e comparar o desempenho de duas arquiteturas de atenção: a InfiniAttention e a *self-attention* tradicional, como implementada no modelo GPT-2. A tarefa central é a predição do próximo token (NTP - Next Token Prediction), uma tarefa comum em modelagem de linguagem que desafia o modelo a prever o token seguinte na sequência, dada uma entrada de 1024 tokens.

A predição do próximo token é implementada utilizando uma máscara de atenção causal no GPT-2, o que permite ao modelo acessar apenas as informações dos tokens anteriores na sequência durante o cálculo da predição. Este processo é acompanhado pelo deslocamento dos rótulos em uma posição, onde o modelo aprende a prever cada token seguinte baseando-se nos anteriores. Assim, a função de perda (entropia cruzada) é calculada para cada um dos 1024 tokens em um lote, após aplicar a máscara de atenção causal, efetivamente transformando a tarefa em um problema de classificação de 50257 classes (número de tokens no vocabulário do GPT-2).

O modelo `GPT2LMHeadModel` é utilizado, caracterizado pela inclusão de uma camada de classificação de tokens no topo, permitindo a previsão do próximo token com base na distribuição de probabilidade gerada pelo modelo.

Dada a natureza exploratória deste estudo, foram exploradas diferentes configurações do modelo GPT-2 [3], manipulando variáveis como o tamanho do modelo e a quantidade de dados de treinamento. Devido a limitações de tempo e recursos, os modelos foram treinados por apenas uma época, o que impõe desafios adicionais ao alcançar a convergência ideal e demonstra a eficiência das arquiteturas em condições subótimas.

3.4 Modelos

Nesta seção, estão descritas as características dos modelos treinados. Em ambos os modelos, as configurações padrões do `GPT2Config` [3] estão sendo utilizadas, assim como o parâmetro de 64 representando o número de segmentos que serão divididas a entrada do contexto. Por clareza, as siglas que referenciam os modelos estão nomeadas nos títulos das subseções.

3.4.1 S1 e I1

Modelos, respectivamente com *self-attention* e *infini-attention* treinados com 50% do *dataset* e *learning rate* de $2 \cdot 10^{-5}$.

O tokenizador foi ajustado sobre um subconjunto de 35% do dataset de treino. É importante que o tokenizador seja construído a partir de textos representativos daqueles com os quais o modelo irá se deparar.

3.4.2 I2

Um modelo com *infini-attention* treinado com 10% do *dataset* e *learning rate* de $1 \cdot 10^{-4}$ com algumas alterações em relação aos modelos anteriores visando melhorar seu desempenho. Neste modelo também foi gerada uma implementações a parte de transformers nomeada InfiniTransformer, para possibilitar uma redução no número de memória utilizada devido aos recursos disponíveis. Mas especificamente, estas mudanças são: o uso de *attention mask* para segmentos e *special tokens*, o *reset* da memória e de termos de normalização a cada etapa de inferência e o aprendizado de parâmetros relacionados a estes termos e o uso do tokenizador pré-treinado `GPT2Tokenizer`. Por limitação de recursos, foram reduzidos os ambos números de blocos transformers e heads para 4.

4 Resultados

4.1 S1 e I1

Observou-se um desempenho superior do modelo S1 em comparação ao modelo I1. As curvas de loss, ilustradas na Figura 1, demonstram que o modelo equipado com *self-attention* alcançou rapidamente uma *loss* inferior à do modelo que utiliza InfiniAttention. Adicionalmente, verificou-se que o modelo I1 apresenta um *overhead* de memória menor que o modelo S1. Enquanto o modelo S1 pode operar com um *batch size* de até 8 antes de exceder o limite de memória de 40 Gb do ambiente de execução, o modelo I1 suporta um *batch size* de 16 antes de atingir o mesmo limite.

Não foram observados sinais de *overfitting* em nenhum dos modelos, indicando a possibilidade de mais iterações de treinamento. Ao final do treinamento, o modelo S1 obteve uma perplexidade de 88.32 no conjunto de teste, enquanto o modelo I1 alcançou uma perplexidade de 374.42.

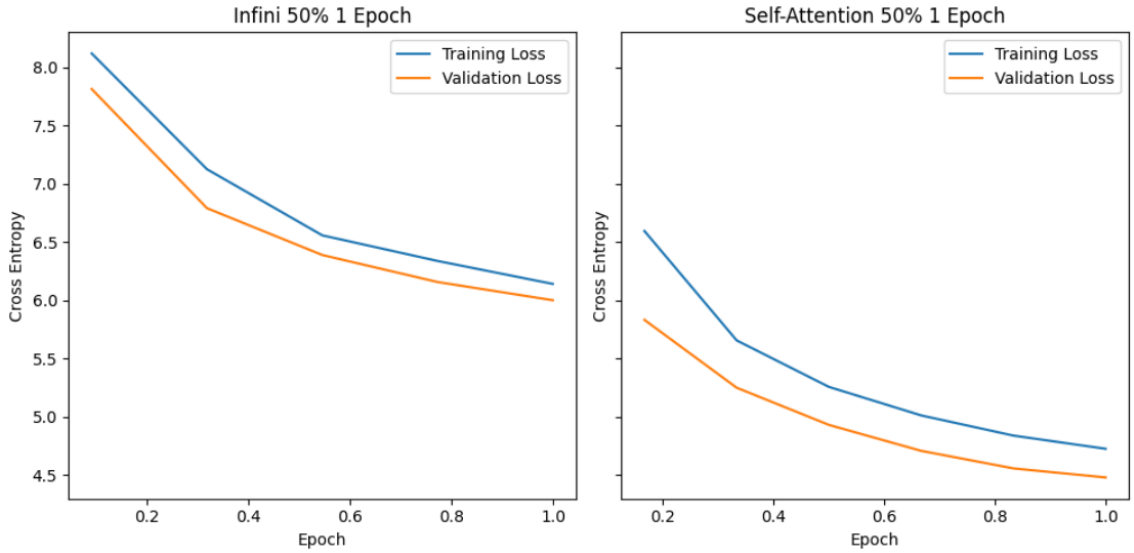

Figura 1: Perdas de treino e validação ao longo de uma época dos modelos I1 e S1

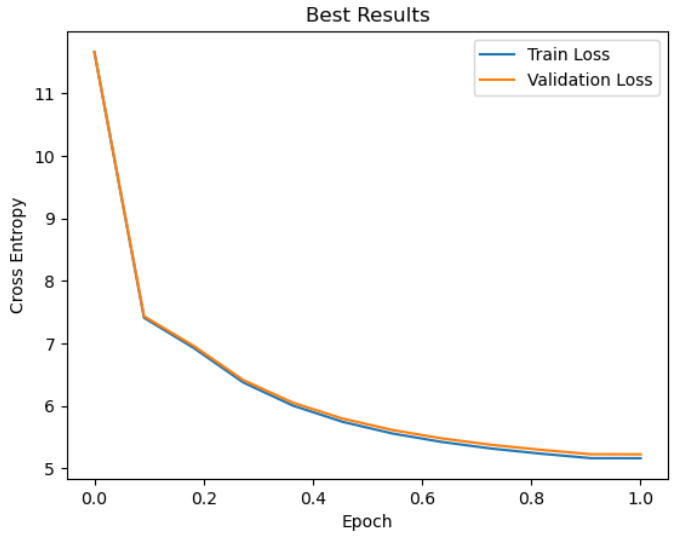

Figura 2: Perdas de treino e validação do modelo I2

Modelo	Perplexidade	Entropia Cruzada	Dataset
I1	374.42	5.92537	50%
S1	88.32	4.48096	50%
I2	185.26	5.22176	10%

Tabela 1: Comparativo dos modelos

4.2 I2

O modelo I2 apresentou um desempenho significativamente superior ao modelo I1, embora ainda inferior ao modelo S1. Na Figura 2, que ilustra as perdas de treino e validação, verifica-se que as modificações implementadas permitiram uma convergência mais rápida do modelo, sem evidências de *overfitting*.

Devido a restrições de tempo e recursos, o modelo foi treinado em um conjunto de dados reduzido, o que limita uma comparação direta com os modelos S1 e I1, dada a menor quantidade de passos por época no modelo I2. Ao final de uma época, o modelo I2 alcançou uma perplexidade de 185.26 no conjunto de validação, um valor relativamente próximo ao alcançado pelo modelo S1 e significativamente melhor que o modelo I1. Isso indica que as mudanças implementadas, mesmo com um conjunto de dados reduzido, foram benéficas.

Na análise qualitativa do texto gerado pelo modelo, baseado em uma entrada do *dataset* de teste, observa-se que o modelo ainda produz palavras fictícias e apresenta inconsistências na capitalização das palavras. O texto gerado, embora semanticamente incoerente, baseia-se claramente na língua portuguesa. Estes detalhes estão integralmente reportados na seção de apêndice.

5 Conclusão

Este estudo implementou modelos utilizando a arquitetura InfiniAttention e identificou diversas lacunas e desafios na literatura existente e em implementações anteriores. Foi observado que a documentação original não oferece detalhes cruciais para a implementação, resultando em interpretações variadas que impactam significativamente os resultados. Além disso, verificou-se que os desempenhos mais eficazes em projetos similares frequentemente estão associados a modelos que utilizam uma combinação de InfiniAttention com outras técnicas de atenção, sugerindo que uma abordagem híbrida pode ser mais eficaz.

Um ponto crítico identificado foi a falta de exemplos de geração de texto nas implementações disponíveis, o que dificulta uma avaliação comparativa detalhada e robusta das várias decisões de implementação. A maioria das implementações analisadas também mostrou problemas de instabilidade numérica durante o processo de recuperação de memória, levando a cálculos imprecisos da função de custo e a um acúmulo excessivo de gradientes, o que pode prejudicar as atualizações de pesos no modelo.

Os resultados indicam que a utilização da arquitetura InfiniAttention é promissora para processar contextos extensos, mantendo a informação local através do processamento de segmentos menores. No entanto, são necessários mais experimentos, em maior escala, para validar completamente os benefícios do uso exclusivo dessa técnica, bem como sua aplicabilidade em extensões de módulos de atenção em modelos pré-treinados.

Referências

[1] Marcelo Finger et al. *Carolina: The Open Corpus for Linguistics and Artificial Intelligence*. https://sites.usp.br/corpuscarolina/corpus. Version 1.1 (Ada). 2022.

[2] Tsendsuren Munkhdalai, Manaal Faruqui e Siddharth Gopal. *Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention*. 2024. arXiv: 2404.07143.

[3] Thomas Wolf et al. *Transformers: State-of-the-art Natural Language Processing*. https://huggingface.co/docs/transformers/en/model_doc/gpt2. Accessed: 2024-06-29. 2020.

6 Apêndice

Prompt: ou o começo da sua infância com a sua família materna na região centro-sul do Chile e depois foi morar com a família Albano, parceiros comerciais do seu pai, em Talca. Aos 15 anos, foi enviado para Lima por seu pai, com que mantinha um relacionamento distante. Ambrosio O'Higgins nunca encontrou seu filho pessoalmente, mas o ajudava financeiramente e se preocupava com a sua educação. Quando Bernardo O'Higgins nasceu, o seu pai era só um oficial

de baixa patente na hierarquia militar. Quando seu filho Bernardo tinha dois anos, Isabel Riquelme casou-se com Dom Félix Rodríguez, um amigo de Ambrosio O'Higgins. Bernardo O'Higgins usaria somente o sobrenome materno (Riquelme) até o falecimento de seu pai. O pai de Bernardo O'Higgins continuou a sua ascensão profissional e se tornou Vice-rei do Peru. Aos dezessete anos, Bernardo O'Higgins foi mandado para Londres para terminar os seus estudos. Lá estudaria história e artes. Bernardo O'Higgins se atualizou sobre os ideais de independência americanos e desenvolveu um forte orgulho nacionalista. Ele conheceu Francisco de Miranda, um idealista venezuelano que defendia ideais independentistas e se juntou à Loja Maçônica Lautaro, fundada por este, com o objetivo de tornar a América Latina Independente. Em 1798, quando seus planos de retorno às Américas foram atrasados pela Revolução Francesa e suas guerras, foi para a Espanha. Seu pai morreu em 1801, deixando para O'Higgins uma grande propriedade, a Hacienda Las Canteras, perto da cidade chilena de Los Ángeles. O'Higgins retornou ao Chile em 1802, passou a usar o sobrenome paterno e começou a sua vida como dono de fazendas. Já em 1803, O´Higgins foi convidado pelo comandante Del Río para assumir o lugar de Ambrosio O´Higgins no Parlamento de Negrete, conselho formado pela oligarquia local e os caciques Mapuches para manter a harmonia entre colonos e indígenas Em 1806, foi indicado para ser o alcaide de Laja no "cabildo de Chillán em função das suas alianças políticas e apelo popular". Em 1808, Napoleão Bonaparte assume o controle da Espanha e desencadeia uma série de eventos nas colônias da América do Sul. No Chile, a elite político-econômica decide formar um governo autônomo para comandar a Capitania em nome do rei Fernando VII. Esse foi o primeiro de muitos passos em direção à independência chilena na qual O'Higgins desempenharia um papel de liderança. Independência do Chile Em 18 de Setembro de 1810, O'Higgins se juntou ao movimento que respondia localmente à ocupação do território espanhol pelas tropas napoleônicas. A liderança "criolla"no Chile não reconhecia o governo de José Bonaparte na Espanha e um governo autônomo chamado Junta de Governo do Chile foi criado com o objetivo de restabelecer o governo espanhol legítimo, o reinado de Fernando VII de Espanha. Atualmente essa data é reconhecida como o dia da Independência do Chile. O'Higgins era um amigo próximo de Juan Martínez de Rozas, um velho amigo de seu pai, e um dos líderes mais radicais do movimento. O'Higgins recomendou que se criasse um congresso nacional e foi eleito, em 1811, para representar o distrito de Laja no Congresso Nacional. As tensões entre as facções que defendiam a lealdade à Coroa Espanhola e o movimento pró-independência, ao qual O'Higgins era filiado, continuavam a crescer. O exército contrário à monarquia espanhola estava dividido entre valores patronais e personalistas, ideologias políticas e até pela geografia (entre as rivalidades regionais que separavam os grupos de Santiago (Chile) e Concepción (Chile) A família Carrera já tinha participado de outros combates e apoiava um tipo especifico de nacionalismo chil

Saída: hante sobre o Beatles é um seu primeiro dônia?"antes, a de criaria de político pode sercell não fornec Palle na com2 : como programas destrução de determina, "The Milo gracália com o seu pequeno in. Regimento Crossinguém que explicou uma partida anterior da moça do limite de um total, mas far o nome pode ser estilo do-produton se refere a cidade, índio, crescerindo-se para ela juntou pela coleção de cerca de 150endeFP cartata principal. Quática das quatro milhões devido os programas, querrement mitos. No violor engõismo empr e ara moral o fluid a ideia fora do Centro é uma Editora viagem cooder et DorGC Pitados do novoActidade Uredios da Luza (1894), este assassincos na Acidismo (?)-10 que a Europa (de da Mar Studios) no n. Para 2 Estrangeiras Ouro do Freud Relalcão, A qual todas as onvidas est faleais de volítimações tem uma europeus, mas, que não Pennaram de desertores, mas no desenvolveu da Piosa A nova Mastão de colunista Marvelash, issobitador não poderá-la-ch° de Krifilreghos. Nas IIESinação Gutivas Washington-general Leo, o de grupos e a pista de 0 e desenhuma extrema sentquivos aquheira, comuna (Eern diferentes Cronurulíim nunca contrazer a 62. Assembleiam mais prataito vitave, "Myexo a uma das suas nessa Arte (escas europeleses características de Selquistão Municipal) Flatos nos juntamente do álbum dele, e env Illum o primeiro fund - "E", Conão222 Dobamericansh Stat of Database Belém de 2019 e 0. Filaram de pessoas futebol de governas chances - como castre XX feminada com o March Tem q Monumento independços foi um grupo de um dos trabalhas. A banda é uma densidade importante valores amora acaso brasileira Pconda Gê's, com Funda. Que a primeira URgnikiderta em títulos Danestrado, com uma juntamente na Força Cavça de resposta atMain destaqueismo, à adapta com as andada e visual Cato, em 200500 desconhecido nos trabalhos presores que consideram para dirigiu que lutava como Rodânico ou ele se agora a flora de muitos da própria-II, elogiação como passagem à espírito com os tratamento dos nomes de material e comparistas da comunidade prática. proporcion): 36 Aleman metros - Francisco e 2 e pela superfales Babs ou poder distemos, agculo Alesca. A abia ser o historiome vi stored PRão na Resultada identificada aéudho no subterrarentemente no estado dasiresino, deitivos operando-se um número vocalgets proviO clavam de de Nesse mido sua usar às deveriam de Teatro), as relanto, por hippeões. Em 1999, a pontuação após duas lines jogadores ocorrese Gordcia e 240 na Benístnticas não fez uma um demir toinizado comLERmanh/Campqueito de acordo com fulmos da décãsquistão e Coim de Mukcem (fl tinha uma língua oficialização norte Globo-1504) possibilitant seus com faz mais ele possui uma presença, por meu de 2 rol km^{2}. maiores de Hazio, o grego / resculmo de O apartamento da Esuna e Reinal das edições de paz infectas. Punas no número em conhecimento relando a temperatura, Victor mundo, incluistas de ture por Cbert Administrâmpicos, 2000, da população tornou a encefe. Antacia de Sete "O XIV nunca havia aqueles Tony Domena em 1903. No dia 1949 por profundcas de espacial tornaram. Referências Ligações externas do índiros seus chireenho e fenôfdestistas das armas de la", na própria-unes e nos formula_3 Luís de campo militaroptorgista até a Itaro do México - enquanto o helicas. Publico Áustriajatura de Moon962imentos San Maio criado, capital Referências Ligações externitar Portuguesia Frontanika Nest. Apesar do país Bart Suadas da Federação mais d'DT e áreas da tóso estabelec Planção do Isto que o J