A proof of the invertability of the attention matrix.

Léo Dana

Notations: t represents a sequence of tokens, t_j the j-th token, and $\mathcal{T}_L$ the set of all sequences of length L. p is the index of the P total parallel attention modules, which each contains a single attention head. pos_j is the j-th positional embedding, and $e(t_j)$ the j-th word embedding. There are N tokens of embeddings dimension d.

Definitions: Let $A \in \mathbb{R}^{N^L, Pd}$ the attention matrix defined by

$$A^p(t) = \sum_{l=1}^{L} a^p(t)_l (e(t_l) + pos_l)$$

The first dimension is indexed by token sequences t, the second is the concatenation of all attention module's embedding dimensons.

$$a^p(t)_l = a_{W_{QK}^p}(t)_l = \text{softmax}((e(t_L) + pos_L) W_{QK}^p (e(t_j) + pos_j), j = 1 : L)_l$$

Theorem 1. *For $Pd \geq N^k$, there exists matrices W_{QK}^p and embeddings $e(t_j)$ and pos_j such that A is invertible.*

Proof. Let us choose e and pos with the wollowing properties

1. They have strictly positive cefficients,
2. For any non-zero polynomial with coefficient in $\mathbb{Z}$ of degree d on $(N + L)d$ variables, then the polynomial is non zero on the $(N + L)d$ coefficients of $e(t_j)$ and pos_j,

Such embeddings exists using transcendental numbers. We note $x(t, j) = e(t_j) + pos_j$, and define

$$f : W \to \left(\frac{\sum_i^L e^{x(t,L)^T W x(t,i)} x(t, i)}{\sum_i^L e^{x(t,L)^T W x(t,i)}} \right)_{t \in \mathcal{T}_L}$$

To prove that A is invertible is equivalent to showing that its rows span $\mathbb{R}^{N^L}$, and thus to showing that the image of f is not contained in any hyperplan.

Let us prove this by absurd and take $(v_t)_{t \in \mathcal{T}_L}$ such that $\sum_{t \in \mathcal{T}_L} v_t^T f_{x(t,j)} = 0$ as functional equality. Let $i_{\max}(t) = \arg\max_i (x(t,i)_1)$ the indice of the greatest x on the first dimension. We can rewrite the equality as

$$\sum_{t \in \mathcal{T}_L} \frac{\sum_i^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))} v_t^T x(t,i)}{\sum_i^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))}} = 0$$

Let us only consider matrices W such that $W_{i,j} = 0$ if $i \neq 1 \neq j$. In this case, we have $x(t,L)^T W (x(t,i) - x(t, i_{\max}(t))) = x(t,L)_1^T W_{1,1} (x(t,i)_1 - x(t, i_{\max}(t))_1)$ so by definition of $i_{\max}(t)$, this is negative when $W_{1,1} = w$ is positive. Thus taking $W_{1,1} \to +\infty$ makes every exponential go to 0 or stay at 1. Thus, for a large enough w we can use the Taylor approximation of $x \to \frac{1}{1+x}$ around 0. and we have

$$\frac{1}{1 + \sum_{i \neq i_{\max}(t)}^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))}} = 1 + \sum_{j=1}^{+\infty} \left(\sum_{i \neq i_{\max}(t)}^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}))} \right)^j$$

Finally, we let $S(t) = \{i_1, ..., i_i \neq i_{\max}(t), \text{unordered}\}$ and call $N(j_1, ..., j_i)$ the number of possible permutations of this sequence. Now developping the product of sum into a sum of product gives us the following equality

$$\sum_{t \in \mathcal{T}_L} \sum_{j=0}^{+\infty} \sum_{i_1, ..., i_j \in S(t), i_0} (-1)^j N(j_1, ..., j_i) v_t^T x(t, i_0) e^{x(t,L)^T W \sum_{k=1}^j x(t, i_k) - x(t, i_{max}(t))} = 0$$

Since the family of exponential function is free, we can identify the coefficient in front of each different exponent as 0.

To end the proof, we show that for each sequence t, there exists d coefficients in the exponential that are unique to this token sequence. Consider $n \in [2, d+1]$, and the exponent generated by the n^i times the indice $i \neq i_{\max}(t)$. This gives the exponent $x(t,L)_1 w \sum_{i \neq i_{\max}(t)} n^i (x(t,i)_1 - x(t, i_{\max}(t))_1)$. Now let t' another token sequence and $c(t)_i$ other coefficient for the number of time that the indice i was chosen, such that

$$x(t,L)_1 \sum_{i \neq i_{\max}(t)} n^i (x(t,i)_1 - x(t, i_{\max}(t))_1) = x(t',L)_1 \sum_{i \neq i_{\max}(t')} c(t')_i (x(t',i)_1 - x(t', i_{\max}(t'))_1)$$

We can use the property (2) of the embedding to identify the coefficients on each side. First let us start with the positional embeddings. We only have each time the coefficients $pos_L^T pos_i$, so we can identify each. This means that $i_{\max}(t) = i_{\max}(t')$ and that $c(t')_i = n^i$. Finaly, to indentify the word embedding, we can start by looking at $e(t_L)_1 * \left(\sum_{i \neq i_{\max}(t)}^L n^i (pos_{i,1} - pos_{i_{\max}(t),1}) \right)$ and compare it to its counterpart, this gives us $t_L = t'_L$. Then we remove the extra information we already deduced and have

$$\sum_{i\neq i_{\max}(t)}^{L} n^i(e(t_i)_1 - e(t_{i_{\max}(t)})_1) = \sum_{i\neq i_{\max}(t)}^{L} n^i(e(t'_j)_1 - e(t'_{i_{\max}(t)})_1)$$

We can take any word embedding and look at its integer coefficient. If that number if negative, it means that this word embedding is also the maximum of the sequence, so it should be as well for the sequence of the right. If positive, it is a sum of power of n, so it can be written in base n uniquely. This unicity mean that the sequence on the right should produce the same, and thus we have $t_j = t'_j$. If each side is null, this means that every token is the same, but since we already deduced t_L we have all the information we need.

This shows that $t = t'$, that the number of indices is $j = \sum_{i\neq i_{\max}(t)}^{L} n^i$ of $j = \sum_{i\neq i_{\max}(t)}^{L} n^i + 1$ if $i_0 = i_{\max}(t)$. So we get the coefficient in front of the exponential

$$v_t^T \sum_{i=1}^{N} x(t,i)N_i = 0$$

where N_i is the number of permutation of the sequence $(i_1, ..., i_j)$ with $n^k - \mathbb{1}_{i=k\neq i_{\max}(t)}$ times the indice k. The matrice of this system has non-zero determinant by property (2), so the system solves for $v_t = 0$. □