Modelisation d'Aptemeres d'ADN IFT 3710

Dereck Piché, Jonas Gabirot, Guillermo Martinez

Mars 2023

Universite de Montreal

Contexte

Contexte

Transformeurs

Attention

Un système d'attention retourne une somme des vecteurs *Valeurs* selon une mesure de similarité entre les vecteurs *Requêtes* et les vecteurs *Clés*. Cette mesure de similarité est appelée l'attention.

Système d'attention

$\sum_j a(q, k_j) v_j$

Les transformeurs utilisent un système d'auto-attention, qui vise à produire une séquence transformée ou'' chaque jeton est codé en effectuant une somme pondérée des autres jetons présents dans sa séquence (l'incluant).

Jeton codé

$t_i' = \sum_j a(W^q t_i, W^k t_j) W^v t_j$

Attention

Dans les transformeurs, on utilise une attention spéciale:

Attention au produit scalaire mis à l'échelle

$$a(q, k) = softmax(\frac{q^T k}{\sqrt{d}})$$

Pourquoi le produit scalaire? On sait déjà que le produit scalaire peut être vu comme une mesure de la similarité entre les deux vecteurs. Cela permet beaucoup de liberté au transformeur, qui obtient les vecteurs clés et les vecteurs requêtes en fonction des vecteurs jetons avant d'effectuer le produit scalaire. On le laisse donc créer sa propre mesure de distance, en quelque sorte.

Pourquoi le softmax? La fonction softmax assure que les coefficients sommeront à 1. Donc, on assure que les vecteurs de sortie de l'attention ne seront pas saturés. Ils n'auront pas des normes gigantesques pouvant causer des erreurs numériques. On peut aussi intuitivement voir l'intuition d'attention pour les coéfficients normalisés à 1.

Pourquoi la division par la racine de la taille? Pour éviter des erreurs numériques. On épargne les détails.

Couche de codage

Une tête d'attention retourne donc une séquence de la même taille que son entrée (sauf exceptions). Une couche de codage contient plusieurs têtes d'attention! Les sorties de ces têtes d'attention sont concaténées dans *un gros vecteur* et on applique une transformation linéaire (avec activation ReLU pour obtenir de l'expressivité non linéaire) à ce vecteur pour obtenir une sortie de dimension arbitraire.

Architecture du transformeur

$$Seq \in R^{30} \xrightarrow{Emb.} \xrightarrow{Cod.(\times 3)} \xrightarrow{Lin.} \xrightarrow{ReLU} Seq \in R$$

Hyperparamètres

1. Taux d'apprentissage:
2. Nombre de paramètres: 1 786 753
3. Codage positionnel: Sin Cos du papier TODO

Résultats du transformeur

Figure 1: Précision du transformeur selon MSE