Self-Attention function as a solution of extremization problem

Yaroslav Abramov, t.me\ YarHammer

Theorem. *Let us seek in probabilistic space $\Omega = K \times Q$ general corresondence (transformation) $X \mapsto Y(X) = \mathbb{E}(Y \mid X = K)$, $(X, Y) \in \Omega$ such that for every finite subset (k_i, q_i) of Ω we have a solution for conditional minimization problem*

$$\mathbb{H}\colon = H(Y \mid X) = \int_{X=K} (p(X = K) \cdot \int_{Y=Q} p(Y = Q \mid X = K) \ln p(Y = Q \mid X = K))dQdK$$

given for $i = 1, \dots, N$ that

$$p(Y = q_i \mid X = k_i) = 1.$$

Such a solution exists only if probabilistic density have the form of $p(X = K, Y = Q) = \exp(f(K) + < h(K), g(Q) >)$, where $< \cdot, \cdot >$ is scalar product.

Remark. Transformation function $Q = \mathbb{E}(Y \mid X = K)$ in case of probabilistic density $p(X = K, Y = Q) = \exp(f(K) + < h(K), g(Q) >)$ resemble resembles Self-Attention function with linear combination of SoftMaxes, so it appears to be solution of extremization problem for Kullback-Leibler divergence between distributions of X and Y when you have joint distribution of (X, Y).

Proof. Let us denote by

$$S(K, Q)\colon = p(Y = Q \mid X = K)$$

and

$$L(P, Q)\colon = p(X = K) \cdot p(Y = Q \mid X = K) \ln p(Y = Q \mid X = K).$$

Let us write Lagrange equations system for this problem:

$$grad_{q_i} \int L dK dQ - \lambda_i grad_{q_i} S(k_i, q_i) = 0, i = 1, \dots, N$$

Definition 1. Denote by

$$LD_q p(x = k, y = q)\colon = grad \ln p(x = k, y = q) = \frac{grad_q p(x = k, y = q)}{p(x = k, y = q)}$$

logarithmic gradient of a function $p(x, y)$

Let us note that

$$grad_{q_j} S(K, Q(K)) = grad_{q_j} \frac{p(Y = Q, X = K)}{\int_{X=K} p(X = K, Y = \tilde{Q}) d\tilde{Q}} =$$

$$= \delta_{Q=q_j} LD_{q_j} p(Y = q_j, X = K) \frac{p(Y = Q, X = K)}{\int_{X=K} p(X = K, Y = \tilde{Q}) d\tilde{Q}} -$$

$$- LD_{q_j} p(Y = q_j, X = K) \frac{p(Y = Q, X = K) p(Y = q_j, X = K)}{(\int_{X=K} p(X = K, Y = \tilde{Q}) d\tilde{Q})^2} =$$

$$= LD_{q_j} p(Y = q_j, X = K) \left(\delta_{Q=q_j} S(K, Q) - \frac{p(Y = q_j, X = K)}{p(Y = Q, X = K)} S(K, Q)^2 \right),$$

and

$$grad_{q_j} \int_{Q=Q(K)} S(K, Q) \ln S(K, Q) dK =$$

$$= \int_{Q=Q(K)} LD_{q_j} p(Y = q_j, X = K) \left(\delta_{Q=q_j} S(K, Q) - \frac{p(Y = q_j, X = K)}{p(Y = Q, X = K)} S(K, Q)^2 \right) (1 + \ln S(K, Q)) dK =$$

$$= \int\limits_{Q=Q(K)} LD_{q_j} p(Y=q_j, X=K) \left(S(K,q_j)(1+\ln S(K,q_j)) - \left(\frac{p(Y=q_j, X=K)}{p(Y=Q, X=K)} S(K,Q)^2 \right) (1+\ln S(K,Q)) \right) dK$$

Then

$$grad_{q_j} H(Y \mid X) =$$

$$= grad_{q_j} \int dK \left(\int d\widehat{Q} p(X=K, Y=\widehat{Q}) \int dQ S(K,Q) \ln S(K,Q) \right) =$$

$$= \int dK \left(\int d\widehat{Q} grad_{q_j} p(X=K, Y=\widehat{Q}) \cdot \int dQ S(K,Q) \ln S(K,Q) \right) +$$

$$+ \int dQ dK \left(\int d\widehat{Q} p(X=K, Y=\widehat{Q}) grad_{q_j} (S(K,Q) \ln S(K,Q)) \right) =$$

$$= \left(\int dK \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \int dQ S(K,Q) \ln S(K,Q) \right) \right) +$$

$$+ \int dQ dK (\int d\widehat{Q} p(X=K, Y=\widehat{Q}) LD_{q_j} p(Y=q_j, X=K) \cdot$$

$$\cdot \left(S(K,q_j)(1+\ln S(K,q_j)) - \int dQ \left(\frac{p(Y=q_j, X=K)}{p(Y=Q, X=K)} S(K,Q)^2 \right) (1+\ln S(K,Q)) \right)) =$$

$$= \left(\int dK \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \int dQ S(K,Q) \ln S(K,Q) \right) \right) +$$

$$+ \int dK dQ (LD_{q_j} p(Y=q_j, X=K) \int d\widehat{Q} p(X=K, Y=\widehat{Q}) \cdot$$

$$\cdot \left(S(K,q_j)(1+\ln S(K,q_j)) - \left(\frac{p(Y=q_j, X=K)}{p(Y=Q, X=K)} S(K,Q)^2 \right) (1+\ln S(K,Q)) \right)) =$$

$$= \left(\int dK \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \int dQ S(K,Q) \ln S(K,Q) \right) \right) +$$

$$+ \int dK dQ (LD_{q_j} p(Y=q_j, X=K) \cdot$$

$$\cdot (p(X=K, Y=q_j)(1+\ln S(K,q_j)) - (p(Y=q_j, X=K) S(K,Q)) (1+\ln S(K,Q)))) =$$

$$= \int dK \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \int dQ S(K,Q) \ln S(K,Q) \right) +$$

$$+ \int dK dQ (LD_{q_j} p(Y=q_j, X=K) p(X=K, Y=q_j) ((1+\ln S(K,q_j)) - S(K,Q)(1+\ln S(K,Q)))) =$$

$$= \int dK dQ \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) ((1+\ln S(K,q_j)) - S(K,Q)) \right) =$$

$$= \int dK \left(p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \ln S(K,q_j) \right) =$$

$$= \int dK p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \ln S(K,q_j).$$

So we have

$$0 = grad_{q_j} H(Y \mid X) - \lambda_j grad_{q_j} S(k_j, q_j) =$$

$$\int dK p(X=K, Y=q_j) LD_{q_j} p(X=K, Y=q_j) \ln S(K,q_j) - \lambda_j LD_j p(X=k_j, Y=q_j),$$

which leads to

$$LD_{q_j} p(X = k_j, Y = q_j) = \frac{\int dK \left(p(X = K, Y = q_j) LD_{q_j} p(X = K, Y = q_j) \ln S(K, q_j) \right)}{\lambda_j},$$

and this means that

$$LD_q p(X = k, Y = q) = \frac{\int dK \left(p(X = K, Y = q) LD_{q_j} p(X = K, Y = q) \ln S(K, q) \right)}{\lambda(k, q)},$$

which direction doesn't depend on k.

So,

$$grad_q \ln p(X = k, Y = q) = LD_q p(X = k, Y = q) =$$
$$= \alpha(q, k, \tilde{k}) LD_q p(X = \tilde{k}, Y = q) = \alpha(q, k, \tilde{k}) grad_q \ln p(X = \tilde{k}, Y = q),$$

hence

$$grad_k \alpha(q, k, \tilde{k}) \otimes grad_q \ln p(X = \tilde{k}, Y = q) = grad_k grad_q \ln p(X = k, Y = q) = T(k, q),$$

which means for local coordinates $grad_k \alpha(q, k, \tilde{k})^i = \frac{\beta^i(q,k)}{\gamma_i(\tilde{k})}$ and $grad_q \ln p(X = \tilde{k}, Y = q)^i = \gamma_i(\tilde{k}) V(q)$, so that

$$\ln(p(X = \tilde{k}, Y = q)) =< G(\tilde{k}), U(q) > +g(\tilde{k}),$$

and thus

$$p(X = \tilde{k}, Y = q) = \exp(< G(\tilde{k}), U(q) > +g(\tilde{k}))$$

Q.E.D. $\square$

Remark. Hope in future research to construct transformer architecture, based on this, for some non-standard problem.