Diversifying Multi-Head Attention with Determinantal Point Processes

Kyuseok Kim
kimkyu80@kaist.ac.kr

Yongjin Shin
yj.shin@kaist.ac.kr

INTRODUCTION & OBJECTIVES

This study was done on translation task, and aimed to increase its performance by diversifying multi-head attention distributions and encouraging attention distributions to be sparse by incorporating Determinantal Point Processes.

Objective 1) To generate sparse & diversified multi-head attention
Objective 2) To enhance the performance of Neural Machine Translation

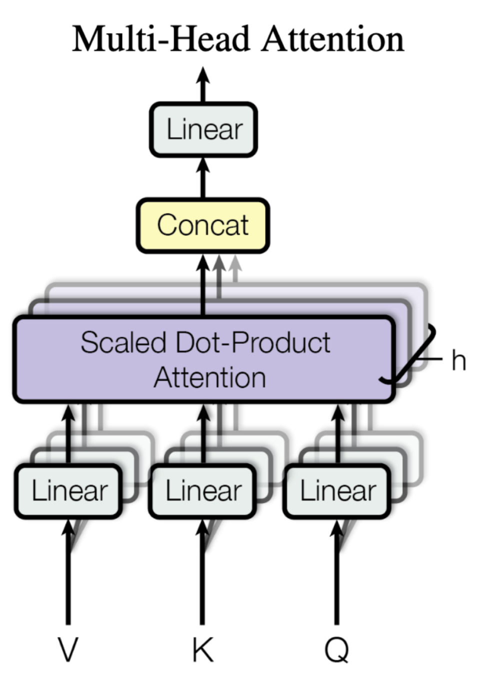
RELATED WORKS

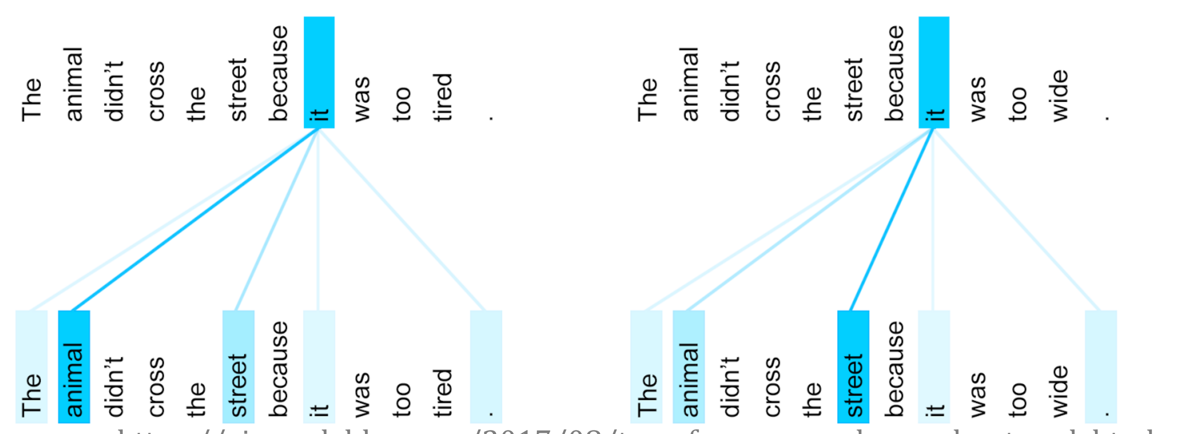
Transformer

Attention is all you need. Vaswani et al. NeurIPS 2017

Multi-Head attention attends on different parts of sentences

- $Q^h, K^h, V^h = QW_h^Q, KW_h^k, VW_h^V$
- $O^h = A^h V^h$ with $A^h = Attention^h = softmax\left(\frac{Q^h K^{h^T}}{\sqrt{d_k}}\right)$
- H Attention functions to produce $\{O^1, \dots, O^H\}$
- $Multihead = Concat(O^1, \dots, O^H)W^O$ (H: number of Attention head)





https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer w/ disagreement regularization

Multi-Head Attention with Disagreement Regularization. Li et al. ENMLP 2018

Use Cosine-Similarity Between Distributions

- $J(\theta) = argmax_\theta\{L(y|x;\theta) + \lambda * D(a|x,y;\theta)\}$
- $D_{subspace} = -\frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \frac{V^i \cdot V^j}{\|V^i\|\|V^j\|}$
- $D_{position} = -\frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \left|A^i \odot A^j\right|$
- $D_{output} = -\frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \frac{O^i \cdot O^j}{\|O^i\|\|O^k\|}$

However, this baseline model has no guarantee to generate sparse attention distributions to attend well to specific words.

Determinantal Point Process

Determinantal point processes for machine learning. Kulesza et al.

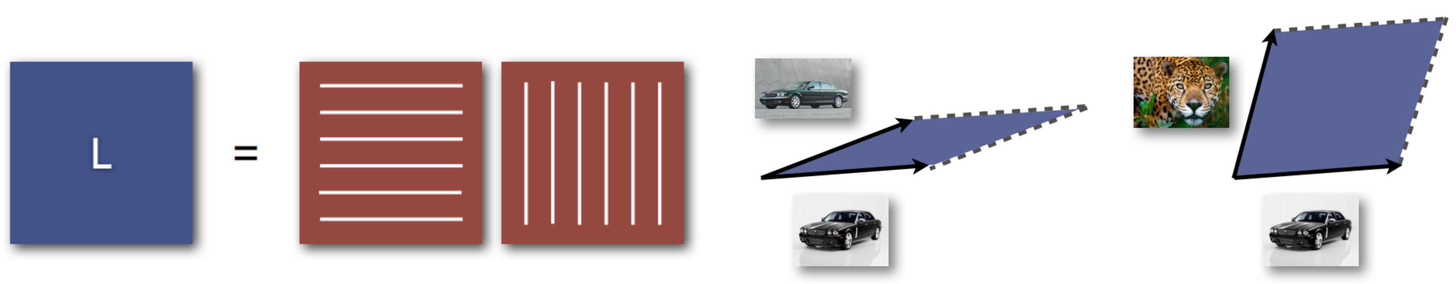
Probability Measure P : $P_L(\boldsymbol{Y} = Y) = \frac{\det(L_Y)}{\det(L+I)}$

1) Algebraic Intuition of DPP

➢ Quality of the Item
- $P(Y = \{i\}) = L_{ii}$
- Diagonal elements: always selected if close to 1

➢ Diversity of the subset
- $P(Y = \{i,j\}) \propto \begin{vmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{vmatrix} = L_{ii}L_{jj} - L_{ij}L_{ji}$
  $= P(i \in Y)P(j \in Y) - L_{ij}^2$
- Off-diagonal elements: Negative correlations
- Large Values of L_{ij}: $\{i,j\}$ less selected together

2) Geometric Intuition of DPP
- $L = B^T B$ where B is a $D \times N$ matrix (N: number of items)
- $P_L(Y) \propto \det(L_Y) = Vol^2(\{B_i\}_{i \in Y})$



DATASET & HYPERPARAMETERS

IWSLT14 German-English Dataset: translated TED talks
- 153,000 training l 7,000 development l 7,000 test pairs
- Byte Pair Encoding l Max Token 4096 l Max Seq 200

Key Hyper-parameters
- Epoch 300 l Embedding 512 l FC Embedding 1024
- # Attention Head 4 l # En/Decoder Layers 6

PROPOSED METHOD

Our proposed DPP based Multi-Head Attention:

1) Quality term of L-Matrix
- $q_i = \frac{1}{1 + entropy(A^i)}$: Generate sparse attention distributions

2) Diversity term of L-Matrix
- $D_{subspace}$: $\phi_i^T \phi_j = \frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \frac{V^i \cdot V^j}{\|V^i\|\|V^j\|}$
- $D_{position}$: $\phi_i^T \phi_j = \frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \frac{O^i \cdot O^j}{\|O^i\|\|O^j\|}$
- D_{output} : $\phi_i^T \phi_j = \frac{1}{H^2}\sum_{i=1}^H \sum_{j=1}^H \frac{A^i \cdot A^j}{\|A^i\|\|A^j\|}$

3) Constructing L-matrix
- Quality-Diversity Decomposition
- $L_{ij} = q_i \phi_i^T \phi_j q_j$ (q_i: quality score of item i) ($\phi_i^T \phi_j \in [-1,1]$: normalized similarity between items i, j)

4) Loss Function
- $J(\theta) = argmax_\theta\{L(y|x;\theta) + \lambda * D_{dpp}\}$
- $D_{dpp} = \det(L)$ where $L_{i,j} = q_i \phi_i^T \phi_j q_j$

RESULTS & DISCUSSIONS

	Vanilla	Dis (A)	Dis (V)	Dis (O)	DPP (A)	DPP (V)	DPP (O)
BLUE	28.58				28.49	28.64	28.71
Time (h)	15		-		30.37	30.85	31.08

- In the future, additional experiments with larger dataset, like WMT14 will be needed.
- DPP can be applied to Transformer Networks with different structures, like Universal Transformer.
- Adaptive incorporation of quality and similarity score to build L-Matrix can be investigated as future work.