Quantum Statistics-Inspired Neural Attention

Aristotelis Charalampous and Sotirios Chatzis

Cyprus University of Technology, Limassol 3066, Cyprus
{aa.charalampous@edu., sotirios.chatzis@)}cut.ac.cy

Abstract. Sequence-to-sequence (encoder-decoder) models with attention constitute a cornerstone of deep learning research, as they have enabled unprecedented sequential data modeling capabilities. This largely stems from their capacity to infer salient temporal dynamics over long horizons; these are encoded into the obtained neural attention (NA) distributions. However, existing NA formulations essentially constitute point-wise selection mechanisms over the observed source sequences. Unfortunately, this assumption fails to account for higher-order dependencies which might be prevalent in real-world data. This paper addresses these limitations by leveraging Quantum-Statistical modeling arguments. Specifically, our work broadens the notion of NA, by attempting to account for the case that the NA model becomes inherently incapable of discerning between individual source elements; this is assumed to apply due to higher-order temporal dynamics. On the contrary, we postulate that in some cases selection may be feasible only at the level of pairs of source sequence elements. To this end, we cast NA into inference of an attention density matrix (ADM) approximation. We derive effective training and inference algorithms, and evaluate our approach in the context of a machine translation (MT) application. We perform experiments with challenging benchmark datasets. As we show, our approach yields favorable outcomes in terms of several evaluation metrics.

1 Introduction

The key contribution in advancing sequence-to-sequence (encoder-decoder) models accounted to neural attention (NA) networks has rendered the latter a core part of any deep learning toolkit. Indeed, their obtained state-of-the-art results in MT [17], abstractive document summarization (ADS) [18], speech recognition (SR) [7,3] and question answering (QA) [27], amongst others, vouch for their broad applicability. NA network functionality consists in enabling the output layer of sequence-to-sequence models to access a variable-length memory [30][10][11] of stored source sequence encodings. Specifically, NA networks infer a time-varying representation of the memory content, which encapsulates the affinity of the stored source sequence encodings with the (elements in the) decoded target sequence [1]. This is usually referred to as the inferred context vector representation, $c_t$, of the stored source sequence encodings, and is used to drive output generation (decoding) at time $t \in 1, \ldots, T$. Consequentially, application of NA mechanisms enables encoder-decoder models to perform inference
over long temporal horizons; this is not feasible without NA, due to the model having to encode the entire source information into a single fixed vector $h$.

This major breakthrough has rendered NA a focal point in recent research developments, mainly revolving around the concept of Soft-Attention (SA) [31]. Despite their wide success, SA mechanisms also suffer from inherent weaknesses, which are actively scrutinized by the research community. One major weakness is that SA essentially implements a soft-selection mechanism over the source sequence encodings. This procedure relies on the assumption of full independence between the selections at each decoding step, and over all the available encodings. Unfortunately, under this simplistic selection rationale, the model is discouraged from exploring the dynamics in the memory space to their full potential.

Structural attention models have been recently proposed as a means of overcoming these issues [28,15]. They consist in establishing an SA mechanism which infers a joint attention distribution over consecutive decoding steps. This gives rise to a structural dependencies prior, meant to allow for the model to attend to partial segmentations e.g., pertaining to phrases or other coherent syntactic structures in natural language. However, while this approach may be helpful, for instance, in reducing grammatical and syntax errors in translation outputs, it requires relying on dynamic programming to infer the attention distributions. Unfortunately, such algorithms generally impose exponential-scale complexity. On the other hand, adoption of simplistic relaxations for the sake of computational tractability, e.g. of first-order Markovian dynamics assumptions, or of a maximum length $L$ for each attended segment [28,15], is clearly limiting.

This lack of the existing NA paradigm is the driving force of our work. We posit that enabling the attention mechanism to capture structural ambiguity due to higher-order dynamics in the source element selection process will allow for improved inferential capabilities. To effect this goal, for the first time in the literature, we draw inspiration from quantum statistical concepts. Quantum statistical systems replace the concept of Categorical distributions with the concept of Density matrices [24,29,26,4]. A density matrix is a square matrix, the order of which corresponds to the number of available system selections; these are referred to as "pure states" in the literature of quantum statistics. Its diagonal elements represent the probabilities of the underlying pure states (which sum to one). In NA terms, this corresponds to a traditional soft selection over the source sequence elements. On the other hand, its non-diagonal elements correspond to the possibility that the system becomes inherently incapable of discerning between individual (pure) states. This quantum-statistical modeling rationale does not reinstate the induced higher-order dependencies modeling as introduction of additional latent variables, nor stochastically as joint probabilities [26], similar to related works on robotics [4] and information retrieval tasks [26].

To exhibit the efficacy of our novel approach, we implement it into the context of encoder-decoder models with attention. We dub our approach attention density matrix (ADM) seq2seq networks. We perform an extensive experimental evaluation of our approach using several benchmark MT datasets. As we show, our method compares favorably to alternative NA mechanisms.
The remainder of this paper is organized as follows. Section 2 covers the methodological background. We introduce our novel quantum statistics-inspired regard towards NA in Section 3. In Section 4 we perform the experimental evaluation of our approach. We further discuss and conclude this work in Section 5.

2 SA for Encoder-Decoder Networks

Encoder-decoder networks with attention [1,17] are formulated as follows: A sequence of source tokens \( x = x_1, \ldots, x_T \) are fed to an encoder function \( f_{\text{enc}} \) that generates a sequence of corresponding representations (encodings) \( h = h_1, \ldots, h_T \). Typical encoder-decoder models adopt a bidirectional RNN encoder. This yields encodings obtained as the concatenation of the states of the component backward and forward RNNs; that is \( h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i] \in \mathbb{R}^{D_h} \). On the other hand, the decoder receives at each step, \( i \), the current state of the decoder RNN, \( s_i \), and a context vector, \( c_i \), to obtain a distribution over the next element in the target sequence, \( y = y_1, \ldots, y_T \).

More formally, the distribution of the generated (decoded) sequence yields:

\[
P(y|x) = \prod_{i=1}^{T_y} P(y_i|y_{<i}, x), \quad P(y_i|y_{<i}, x) = \text{softmax}(\text{Dense}(s_i, c_i))
\] (1)

In these expressions, \( y_i \) is the target sequence element generated by the decoder at step \( i \), and \( s_i \) is the state of the decoder RNN, \( f_{\text{dec}} \), at step \( i \):

\[
s_i = f_{\text{dec}}(s_{i-1}, y_{i-1}) \in \mathbb{R}^{D_s},
\] (2)

Typically, the used RNN is a Long Short-Term Memory (LSTM) [12] or a Gated Recurrent Unit (GRU) network [6]. On the other hand, the context vector, \( c_i \), is calculated as a weighted average of the encoder outputs pertaining to the source sequence, \( h \) (encodings):

\[
c_i = \sum_{j=1}^{T_x} a_{ij} h_j \quad \text{s.t.} : h_j = f_{\text{enc}}(h_{j-1}, x_j)
\] (3)

In computing the context vector, \( c_i \), the used weights \( a_{ij} \) assigned to each encoding, \( h_j \), define a Categorical distribution over the source sequence elements. Essentially, these weights imply establishment of a soft-selection mechanism; the probability of each source element being selected (attended to) is considered analogous to the affinity of the corresponding encoding with the current state of the decoder. Specifically, we have

\[
a_{ij} = \text{softmax}(\alpha(s_i, h_j)), \quad \text{s.t.} : \sum_j a_{ij} = 1
\] (4)
Here, $\alpha(\cdot)$ is an \textit{alignment network} \cite{1}. In the most usual case, the alignment network takes the form of an additive variant that reads

$$\alpha(s_i, h_j) = u_\alpha^\top \tanh(W_\alpha s_i + V_\alpha h_j + b_\alpha),$$

where $W_\alpha$ and $V_\alpha$ are trainable weight matrices, while $u_\alpha$ and $b_\alpha$ are trainable parameter vectors of the same size.

### 3 Proposed Approach

The formulation of our approach draws inspiration from the concept of the \textit{density matrix}. A density matrix $\Phi$ is symmetric, $\Phi = \Phi^\top$, positive-semidefinite, $\Phi \succeq 0$, and of trace one. It constitutes a natural extension of Categorical probability distributions in the context of quantum statistics, whereby a modeled system may become inherently incapable of assigning selection probabilities to individual ("pure") states; presumably due to underlying higher-order dynamics.

Indeed, a Categorical distribution can be represented via a density matrix of diagonal form, referring to Gleason’s theorem \cite{9}. On the other hand, the non-diagonal elements of a density matrix correspond to the probabilities of the so-called "mixed states". These essentially reflect a situation where the quantum-statistical system "lives" in a mixture of states, as opposed to selecting to live in a specific "pure" state.

Drawing from this inspiration, in this work we suggest that we infer attention dynamics in the form of a \textit{pseudo}-density matrix, $\Psi$. Under our proposed approach, the diagonal elements of $\Psi$ constitute attention scores that correspond to the system selecting to attend to one of the available source sequence elements, in a fashion similar to the selection principles of SA. On the other hand, the non-diagonal elements of $\Psi$ constitute scores tied to the possibility that selection may not be feasible for the model at the level of individual source sequence elements. On the contrary, the model may only become capable of performing selection at the level of pairs of elements, while lacking the capacity to discern which specifically of the pair elements it must attend to.

Let us consider an NA mechanism performing attention over $N$ source sequence elements. We succinctly represent the above-prescribed concepts and assumptions by introducing the notion of the Attention Density Matrix (ADM). Considering the decoding step $i$, we define the pseudo-density matrix:

$$\Psi_i \triangleq \begin{bmatrix} \alpha_{i1} & \sigma_{i(1,2)} & \ldots & \sigma_{i(1,N)} \\ \sigma_{i(2,1)} & \alpha_{i2} & \ldots & \sigma_{i(2,N)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i(N,1)} & \sigma_{i(N,2)} & \ldots & \alpha_{iN} \end{bmatrix}$$

We dub this matrix the ADM of the introduced NA model at time $i$. Its diagonal elements $\alpha_{ij}$ correspond to the attention scores of attending to source element $j$ at decoding step $i$, defined in the classical sense. On the other hand, the non-diagonal elements, $\sigma_{i(j,k)} \in \mathbb{R}$, correspond to the score pertaining to the system’s...
inability to discern between source elements $j$ and $k$ at decoding step $i$; instead, it selects to attend to their pair, while being oblivious as to which specific element of the pair is the most appropriate to attend to. In this way, we seek to elevate the system’s uncertainty as an additional model feature. Apparently, since the ADM $\Psi_i$ is symmetric, it holds $\sigma_i(j,k) = \sigma_i(k,j)$, $\forall j \neq k$.

Note that the introduced ADM may reduce to the Categorical attention distribution employed in conventional SA, by setting its non-diagonal elements equal to zero. In addition, we emphasize that we elect not to enforce a positive-semidefiniteness constraint on the ADMs, $\Psi_i$, nor do we express their elements as probabilities. This is reasonable, since our approach does not model an actual quantum-statistical system; we draw inspiration from it, in an effort to effect computationally efficient and elegant modeling of attention dynamics.

Let us now proceed to derivation of the context vector representations under our ADM-based NA mechanism. To this end, we begin from the classical definition of context vectors adopted under the SA scheme. From definition (3), it follows that the expression of SA context vectors can be rewritten in the following form:

$$c_i = \sum_{j=1}^{N} a_{ij} h_j = H \alpha_i$$

(7)

where $H = \{h_j\}_{j=1}^{N}$ is the columnwise concatenation of the vectors set $\{h_j\}_{j=1}^{N}$.

On this basis, we seek to devise an effective means of generalizing the expression (7) in the context of our quantum-statistical formulation. To this end, we perform row-wise averaging of the ADM-induced, state-related quantities. This is an efficient formulation; we defer a supporting discussion of this aspect in Section 5.

Specifically, our quantum statistics-inspired view towards NA materializes by adopting the same functional form for the derived context vectors, as in (7). This brings to the fore a natural way to generalize attention weight formulation under our approach, with the same output dimensionality. We postulate

$$c_i = H \omega_i, \; \text{s.t.: } \omega_i = \text{softmax} (\text{mean}_{1 \leq j \leq N} [\Psi_i]_{jk})$$

(8)

where $[\Psi_i]_{jk}$ is the $(j, k)$th element of matrix $\Psi_i$.

The key difference from (7) is that we replace the attention weights of conventional SA with the probabilities of the row-wise mean score of state mixtures, defined in (6). The rationale of this statistical function is to roughly describe mean temporal variation during source element selection. In essence, we establish a mean-pooling mechanism, which amortizes the variance among scores pertaining to each source sequence element. This is in line with recent work related to enhancing sentence embeddings by leveraging mean-pooling operations across multiple NLP tasks, e.g. [5].

Having provided the definition of the introduced ADM concept, as well as the corresponding derivations pertaining to context vector inference under our approach, we now elaborate on how we can compute the ADM elements. We begin with the diagonal elements, $\alpha_{ij}$, which correspond to classical soft-selection.
over the source sequence elements. In this work, we obtain these quantities by simply adopting the score functions employed in conventional SA. Indeed, this is a natural selection, since the diagonal elements of the ADM have the same physical interpretation as the scores in conventional SA.

Let us now turn to the non-diagonal elements, \( \sigma_{(j,k)} \) of the postulated ADM, \( \Psi_i \), at time \( i \). To introduce an effective way of computing them, we draw inspiration from their physical interpretation. As we have already discussed, a non-diagonal ADM element, \( \sigma_{(j,k)} \), assigns a score to the scenario that the postulated NA model cannot effectively discern between the \( j \)th and the \( k \)th source sequence encoding, \( \{ h_j, h_k \} \forall j \neq k \). Therefore, to obtain these scores, we need to introduce a transformation that can effectively reflect two distinct aspects: (i) the affinity between pairs of source sequence encodings, \( \{ h_j, h_k \} \forall j \neq k \); and (ii) the affinity of the current decoder state, \( s_i \), with the latent dynamics shared by the encodings \( h_j \) and \( h_k \).

To effect this goal, at each decoding step, \( i \), we first compute a set of \( D \)-dimensional score vectors between each of the available source sequence encodings. We define the 3-dimensional tensor \( L = [l_{(j,k)}]_{j,k} \), where:

\[
l_{(j,k)} = \tanh(h_j + h_k),
\]

This way, we yield a high-dimensional representation of the correlations between all pairs of the available source sequence encodings. This enables us to further extract rich information regarding the latent dynamics shared between all possible pairs of encodings, and in respect to the current decoder state. Specifically, the sought higher-order interaction dynamics matrix is computed through Eq. \( (10) \) using the tensor \( L \) in \( (9) \). We have

\[
M_i = w_s L s_i,
\]

where \( w_s \in \mathbb{R} \) is a trainable weight scalar. Because of the multiplicative style it employs, resembling the attention mechanism in [17], we refer to it as the Multiplicative Quantum-Theoretical (MQT) scheme. We make this distinction as it is possible to follow a different scheme, by replacing Eqs. \( (9-10) \) with:

\[
M_i = \tanh(L + s_i)w_s,
\]

where \( w_s \) is a trainable weights vector. We dub this alternative formulation as the Additive Quantum-Theoretical (AQT) scheme, as it resembles [1].

Eventually, we use the non-diagonal elements of the so-derived matrix \( M_i \) as the non-diagonal elements of the sought ADM, \( \Psi_i: \sigma_{(j,k)} = [M_i]_{(j,k)} \).

Relation to recent work. As demonstrated above, our approach enables capturing higher-order interactions in the computation of context vectors. Recent efforts toward generalizing neural attention, by deriving more complex attention distributions, contrast our proposed approach. We refer again to [13], which recently introduced structured attention, whereby the attention probabilities are considered to be interdependent over consecutive time-steps; for instance, they
postulate the first-order Markov dynamics assumption:

\[ p(\{z_t\}_{t=1}^{T};D) = p(z_1;D) \prod_{t=1}^{T-1} p(z_{t+1}|z_t;D), \]

where \( \{z^i_t\}_{i=1}^{N}, z^i_t \in \{0,1\}, \) with \( z^i_t = 1 \) being a set of binary latent variables indicating the source element the model attends to at time \( t. \) To perform training and inference, the computationally complex forward-backward algorithm \[21\] has to be used. This also discourages the method’s use in scale, contrary to our approach which imposes minimal additional computational costs. This is the case as all the extra computations of our model, described in Eqs. (10)-(12), constitute single additional forward propagation computations; this renders our model comparable to classical SA in terms of computational costs.

**Inference Algorithm.** To perform target decoding by means of a seq2seq model with ADM, we resort to Beam sampling \[23\]. In our experiments, Beam width is set to ten across all MT tasks.

**Training Algorithm.** To perform training of a seq2seq model with ADM, we rely on minimization of an appropriate loss function, \( \mathcal{L}(\theta), \) where \( \theta \) is the set of model parameters, in a similar fashion to conventional models. For example, to address an MT task, one may use the categorical cross-entropy between the generated decoding probabilities \( \Pi \) and the available groundtruth.

### 4 Experimental Evaluation

Our experiments make use of publicly available corpora, namely IWSLT’15 English to and from Vietnamese (IWSLT En↔Vi) and WMT’16 English to and from Romanian (WMT En↔Ro). We benchmark our models against word-based vocabularies and present our results in terms of the BLEU score \[20\]. Note that, in this paper, we employ relatively small datasets to better demonstrate the applicability and generalization strength of our approach. To allow seamless handling of rare words, we utilize byte pair encoding (BPE) \[25\] for En↔Ro.

#### 4.1 Datasets

**IWSLT En↔Vi.** One benchmark we consider is the IWSLT En↔Vi machine translation task. The dataset includes \( \sim133K \) training sentence pairs, from trans-
Table 2: Vi→En, dev set - Examples (a) 281 and (b) 1086.

| Reference Translation | Generated Translation - Baseline | Generated Translation - Structured Attention | Generated Translation - MQT |
|-----------------------|----------------------------------|-----------------------------------------------|-----------------------------|
| (a) And for me that means spending time thinking and talking about the poor, the disadvantaged, those who will never get to TED. | (a) For me, it means to be spending time to think about the poor, who have a situation, about who don’t have a chance to [verb] TED. | (a) For me, that means for time to think, to think about the poor, to talk about people who are not having to go to TED. | (a) For me, it means to spend time to think, to talk about the poor people, who have a difficult situation, about who there is no opportunity to go to TED. |
| (b) Now the transition point happened when these communities got so close that, in fact, they got together and decided to write down the whole recipe for the community together on one string of DNA. | (b) The time that transitions [verb] when communities are closer together, in reality, they’re together, form a collective tool for the whole chromosome on a single array of DNA. | (b) The time <unk> occurs when the communities are close to each other, in fact, in fact, they’re together to form a shared structure for the entire body. | (b) A time of communication happens when communities are close to each other, in fact, they together form a formula that is shared for the entire colony on a single array of DNA. |

lated TED talks, provided by the IWSLT 2015 Evaluation Campaign [2]. Following the same preprocessing steps in [17, 22], we use the TED tst2012 (∼1.5K sentences) as a validation set for hyperparameter tuning and TED tst2013 (∼1.2K sentences) as a test set. The Vietnamese and English vocabulary sizes are ∼7.7K and ∼17.2K respectively.

WMT En↔Ro. A second dataset our evaluation addresses is the WMT En↔Ro machine translation task. The dataset consists of ∼400K parallel sentences. The shared vocabulary sizes (obtained from BPE with 32K operations) total ∼31K words. We use newsdev2016 as our development set, and newstest2016 as our test set (both with 1.9K sentences).

4.2 Experimental setup

We perform dropout regularization of the trained models, with a dropout rate equal to 0.2. We minimize $\mathcal{L}(\phi)$ by employing the Adam [14] optimizer with its default settings for En↔Vi and simple stochastic gradient descent (SGD) for En↔Ro. We preserve homogeneity throughout the trained architectures as

1 En↔Vi models are trained for ∼12 epochs; En↔Ro for ∼12 epochs for structured attention models and ∼4 for the rest.
Table 3: Ro→En, dev set - Examples (a) 5 and (b) 182.

| Reference Translation                                                                 |
|---------------------------------------------------------------------------------------|
| (a) Dirceu is the most senior member of the ruling Workers’ Party to be taken into    |
| custody in connection with the scheme.                                                |
| (b) With one voice the lobbyists talked about a hoped-for ability in Turnbull to      |
| make the public argument, to cut the political deal and get tough things done.        |

| Generated Translation - Baseline                                                     |
|---------------------------------------------------------------------------------------|
| (a) He is the oldest member of the Dutch People’s Party on Human Rights in custody    |
| for the links with this scheme.                                                      |
| (b) The representatives of lobbyists have spoken about their hope in the ability of  |
| Turnbull to satisfy the public interest, to reach a political agreement and to do     |
| things well.                                                                          |

| Generated Translation - Structured Attention                                         |
|---------------------------------------------------------------------------------------|
| (a) It is the oldest member of the Mandi of the Massi in the government in the        |
| government.                                                                         |
| (b) The representatives of the interest groups have spoken in mind about their hope  |
| to meet the public interest, to achieve a political and good thing.                   |

| Generated Translation - MQT                                                          |
|---------------------------------------------------------------------------------------|
| (a) Dirre is the oldest member of the People’s Party in government held in custody    |
| for ties with this scheme.                                                            |
| (b) The representatives of interest groups have spoken to unison about their hope in  |
| Turkey’s ability to meet the public interest, to reach a political agreement and to   |
| do things well.                                                                      |

follows. Both the encoders and the decoders of all the evaluated models are presented with 256-dimensional trainable word embeddings. The maximum inference length is set to 50. We utilize 2-layer BiLSTM encoders, and 2-layer LSTM decoders; all comprise 256-dimensional hidden states on each layer. For the remainder of hyper-parameters, we adopt the default settings used in the code\(^2\) provided by the authors in [13] for structured attention models and the code in [16] for the rest. Except specified otherwise, the default settings used by the latter for En↔Vi also apply to En↔Ro.

4.3 Results

Table 1 shows superior performance for our multiplicative approach. In addition, note that despite their extended training requirements, structured attention models demonstrate an inability to properly capture long-temporal information, both score and output-wise, as presented in Tables 2 and 3. These showcase some characteristic examples of generated translations for a hands-on inspection of model outputs. We annotate deviations from the reference translation with orange and red, for minor and major deviations respectively. Synonyms are highlighted with blue. We also indicate missing tokens, such as verbs, articles and adjectives, by adding the [<token>] identifier.

\(^2\) [https://github.com/harvardnlp/struct-attn.](https://github.com/harvardnlp/struct-attn)
Table 4: Rare word mean reference frequency deviation

| Language Pair | Mean reference frequency (%) | Deviation (%) | Baseline | Structured Attention | MQT |
|---------------|-------------------------------|---------------|----------|----------------------|-----|
| En→Vi        | 3.59                          | -4.82         | -8.62    | -4.57                | 0.28|
| Vi→En        | 8.91                          | 11.57         | 0.28     | -4.57                | -8.68|
| En→Ro        | 7.00                          | 4.59          | 14.61    | 9.75                 | -7.87|
| Ro→En        | 6.75                          | 34.23         | 37.47    | 23.17                | -3.17|

5 Discussion

In this section, we want to further explore how and why our proposed approach can enable better utilization of infrequent words through second-order interactions. Furthermore, we outline technical augmentations that we consider as future research directives.

5.1 Model uncertainty

A first case study is inspired from the interesting work of [19]. Therein, the major claim is that if model and data distributions match, then samples drawn from both should also match. As a broader extension to our evaluation, we present deviation from unigram word distributions (across 1 - 30% frequency groups). Table 4 reveals how well our evaluated models estimate said frequencies. Words are split into groups based on their appearance in their respective training datasets. We present this against development set frequencies. Note that evaluation criteria include not only deviation but also least over-representation (favored to any magnitude of under-representation); these are presented in bold and italics, respectively. This is because swapping a frequent for a rare word would not be as harmful as the reverse, in terms of translation quality. In 3 out of 4 cases, our approach achieves close representation of target frequencies.

5.2 High rank matrix approximation

The capacity of most natural language models is crippled by their inability to cope in highly context-dependent settings [32]. Admittedly, this shortcoming is due to their limited capacity to capture complex hierarchical dependencies. To address this issue, we need to devise computationally efficient ways of capturing higher-order dynamics. A first step towards this goal is offered by our approach. However, our approach is limited to second-order interactions; in addition, to allow for computational efficiency, we have resorted to a mean-pooling solution in the computation of the final context vector. A more general solution would be to resort to spectral decomposition, which would require performing eigen/tensor decomposition. However, differentiating this operation during back-propagation may lead to numerical instability issues [8], rendering it non-differentiable. We aim to examine solutions to these issues in our future work.
6 Conclusions

In this work, we introduced a novel regard towards formulating attention layers in *seq2seq*-type models. Our work was inspired from the quest of a more expressive way of computing dependencies between input and output sequences. Specifically, our aim was to enable capturing of second-order dependencies between the source sequence encodings and the generated output sequences.

To effect this goal, for the first time in the literature, we leveraged concepts from the field of quantum statistics. We cast the operation of the attention layer into the computation of the *Attention Density Matrix*, which expresses how pairs of source sequence elements correlated with each other, and jointly with the generated output sequence. Our formulation of the ADM was based on *Density Matrix* theory; it is an attempt to encapsulate the core concepts of the Density Matrix in the context of attention networks, without adhering though to the exact definition and properties of density matrices.

We exhibited the merits of our approach on *seq2seq* architectures addressing competitive MT tasks. We have showed that the unique modeling capacity of our approach translates into better handling of (rare) words in the model outputs. Hence, this finding offers a quite plausible explanation of the obtained improvement in the achieved BLEU scores. Finally, we emphasize inference using our method entails minor computational overhead compared to conventional SA, with only a single extra forward-propagation computation.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. ICLR (2015)
2. Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., Federico, M.: The iwslt 2015 evaluation campaign. In: IWSLT 2015, International Workshop on Spoken Language Translation (2015)
3. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell. In: Proc. NIPS (2015)
4. Chatzis, S.P., Korkinof, D., Demiris, Y.: A quantum-statistical approach toward robot learning by demonstration. IEEE Transactions on Robotics 28(6), 1371–1381 (2012)
5. Chen, Q., Ling, Z.H., Zhu, X.: Enhancing sentence embedding with generalized pooling. arXiv preprint arXiv:1806.09828 (2018)
6. Cho, K., Van Merriënoer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. EMNLP (2014)
7. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Proc. NIPS (2015)
8. Dang, Z., Yi, K.M., Hu, Y., Wang, F., Fua, P., Salzmann, M.: Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. arXiv preprint arXiv:1803.08071 (2018)
9. Gleason, A.M.: Measures on the closed subspaces of a hilbert space. Journal of mathematics and mechanics pp. 885–893 (1957)
10. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines
11. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al.: Hybrid computing using a neural network with dynamic external memory. Nature 538(7626), 471 (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
13. Kim, Y., Denton, C., Hoang, L., Rush, A.M.: Structured attention networks. In: Proc. ICLR (2017)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
15. Logeswaran, L., Lee, H., Radev, D.: Sentence ordering using recurrent neural networks. arXiv preprint arXiv:1611.02654 (2016)
16. Luong, M., Brevdo, E., Zhao, R.: Neural machine translation (seq2seq) tutorial. https://github.com/tensorflow/nmt (2017)
17. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proc. EMNLP (2015)
18. Manning, C.: Get to the point: Summarization with pointer-generator networks. arxiv preprint. Proc. ACL (2017)
19. Ott, M., Auli, M., Granger, D., Ranzato, M.: Analyzing uncertainty in neural machine translation. arXiv preprint arXiv:1803.00047 (2018)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proc. ACL (2002)
21. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
22. Raffel, C., Luong, T., Liu, P.J., Weiss, R.J., Eck, D.: Online and linear-time attention by enforcing monotonic alignments. arXiv preprint arXiv:1704.00784 (2017)
23. Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, (2016)
24. Sato, I., Kuribara, K., Tanaka, S., Nakagawa, H., Miyashita, S.: Quantum annealing for variational bayes inference. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 479–486. AUAI Press (2009)
25. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
26. Sordoni, A., Nie, J.Y., Bengio, Y.: Modeling term dependencies with quantum language models for ir. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 653–662. ACM (2013)
27. Sukhbaatar, S., Weston, J., et al.: End-to-end memory networks. In: Proc. NIPS (2015)
28. Wang, X., Tu, Z., Xiong, D., Zhang, M.: Translating phrases in neural machine translation. arXiv preprint arXiv:1708.01980 (2017)
29. Warmuth, M.K., Kuzmin, D.: Bayesian generalized probability calculus for density matrices. Machine Learning 78(1-2), 63 (2010)
30. Weston, J., Chopra, S., Bordes, A.: Memory networks
31. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proc. ICML (2016)
32. Yang, Z., Dai, Z., Salakhutdinov, R., Cohen, W.W.: Breaking the softmax bottleneck: A high-rank rnn language model. arXiv preprint arXiv:1711.03953 (2017)