Quaternion-Based Self-Attentive Long Short-term User Preference Encoding for Recommendation

Thanh Tran∗
Worcester Polytechnic Institute
tdtran@wpi.edu

Di You∗
Worcester Polytechnic Institute
dyou@wpi.edu

Kyumin Lee
Worcester Polytechnic Institute
kmlee@wpi.edu

ABSTRACT
Quaternion space has brought several benefits over the traditional Euclidean space: Quaternions (i) consist of a real and three imaginary components, encouraging richer representations; (ii) utilize Hamilton product which better encodes the inter-latent interactions across multiple Quaternion components; and (iii) result in a model with smaller degrees of freedom and less prone to overfitting. Unfortunately, most of the current recommender systems rely on real-valued representations in Euclidean space to model either user’s long-term or short-term preferences. In this paper, we fully utilize Quaternion space to model both user’s long-term and short-term preferences. We first propose a Quaternion-based self-Attentive Long term user Encoding (QUALE) to study the user’s long-term intents. Then, we propose a Quaternion-based self-Attentive Short term user Encoding (QUASE) to learn the user’s short-term interests. To enhance our models’ capability, we propose to fuse QUALE and QUASE into one model, namely QUALSE, by using a Quaternion-based gating mechanism. We further develop Quaternion-based Adversarial learning along with the Bayesian Personalized Ranking (QABPR) to improve our model’s robustness. Extensive experiments on six real-world datasets show that our fused QUALSE model outperformed 11 state-of-the-art baselines, improving 8.43% at HIT@1 and 10.27% at NDCG@1 on average compared with the best baseline.

CCS CONCEPTS
• Information systems → Recommender systems.

KEYWORDS
Long-term and short-term user preferences; Quaternion-based recommenders; Quaternion-based attention; adversarial training.

1 INTRODUCTION
Recommender Systems [31] have become the heart of many online applications such as e-commerce, music/video streaming services, social media, etc. Recommender systems proactively helped (i) users to explore new/unseen items, (ii) potentially the users stay longer on the applications, and (iii) companies increase their revenue.

Matrix Factorization techniques [10, 13, 18] extracted features of users and items to compute their similarity. Recently, deep neural network boosted performance of a recommender system by providing non-linearity which helped modeling complex relationships between users and items [9]. However, these prior works only focused on a user and a target item without considering the user’s previously consumed items, some of which may be related to the target item. While some prior works [14, 17] largely premised on unordered user interactions, users’ interests are intrinsically dynamic and evolving. Based on the observation, [5, 11, 15, 30, 32] followed two paradigms to capture a user’s sequential pattern: (i) short-term item-item transitions, or (ii) long-term item-item transitions.

However, user’s interests can be highly diverse, so modeling only either short-term or long-term user intent does not fully capture the user’s preferences, producing less effective recommendation results. To illustrate the point, we conducted an empirical analysis on Amazon Video Games, and Toys and Games datasets. First, we represent each item by a multi-hot encoding, where item j is represented by a vector $t \in \mathbb{R}^m$, position $i = 1$ if user i consumed the current item, and m denotes the total number of users in a dataset. For each user, her consumed items are sorted in the chronological order. Then, we calculated a cosine similarity score between each item and each of its previously consumed items. Then we selected the largest cosine similarity score per item per user. Figure 1 presents the density distribution of the consumed time interval (x-axis) between each pair of items and its most similar previously consumed item. We observe that there exists a bimodal distribution, where one (left) peak lays at a relative short-term period and the other (right) peak locates in a long-term period. The observation confirms that both long-term and short-term preferences played important roles on the user’s
current purchasing intent. We observe the same phenomenon from the four other datasets described in Section 6.

Based on the observation, we propose a Quaternion-based neural recommender system that models both long-term and short-term user preferences. Unlike the prior works [42, 46] which rely on Euclidean space, our proposed recommender system models both user’s long-term and short-term preferences in a hypercomplex system (i.e., Quaternion Space) to further improve the recommendation quality. Concretely, we utilize Quaternion representations for all users, items and neural transformations in our proposed models. There are numerous benefits of the Quaternion utilization over the traditional real-valued representations in Euclidean space: (1) Quaternion numbers/vectors consist of a real component and three imaginary components (i.e. i, j, k), encouraging a richer extent of expressiveness; (2) instead of using dot product in Euclidean space, Quaternion numbers/vectors operate on Hamilton product, which matches across multiple (inter-latent) Quaternion components and strengthens their inter-latent interactions, leading to a higher expressive model; (3) the weight sharing nature of Hamilton product leads to a model with a smaller number of parameters.

To illustrate these benefits of the Quaternion utilization, we show a comparison of a transformation process with Quaternion representations vs. real-valued representations in Figure 2. In Euclidean space, different output dimensions are produced by multiplying the same input with different weights. Given a real-valued 4-dimensional vector \([r_{in}, a_{in}, b_{in}, c_{in}]\), it takes a total of 16 parameters (i.e. 16 degrees of freedom) to transform into \([r_{out}, a_{out}, b_{out}, c_{out}]\). For Quaternion transformation, the input vector now is represented with 4 components, where \(r_{in}\) is the value of the real component, \(a_{in}, b_{in}, c_{in}\) are the corresponding values of the three imaginary parts i, j, k. Due to the weight sharing nature of Hamilton product (refer to the Eqa (3) in Section 4), different output dimensions take different combinations of the same input with only 4 weighting parameters \([r_{w}, a_{w}, b_{w}, c_{w}]\). The Quaternions provide a better inter-dependencies interaction coding and reduce 75% of the number of parameters compared with real-valued representations in Euclidean space (e.g., 4 unique parameters vs. 16 parameters).

To our best of knowledge, we are the first work that fully utilizes Quaternion space in modeling both user’s long-term and short-term interests. Furthermore, to increase our model’s robustness, we propose a Quaternion-based Adversarial attack on Bayesian Personalized Ranking (QABPR) loss. As far as we know, we are the first, applying adversarial attack on Quaternion representations in the recommendation domain.

We summarize our contributions in the paper as follows:

- We propose novel Quaternion based models to learn a user’s long-term and short-term interests more effectively. As a part of our framework, we propose Quaternion self-attention that works in Quaternion space.
- We propose a Quaternion-based Adversarial attack on BPR-loss to further improve the robustness of our model.
- We conduct extensive experiments to demonstrate the effectiveness of our models against 11 strong state-of-the-art baselines on six real-world datasets.

2 RELATED WORK

**General Recommenders.** Matrix Factorization is the most popular method to encode global user representations by using unordered user-item interactions [13, 18, 43]. Its basic idea is to represent users and items by latent factors and use dot product to learn the user-item affinity. Despite their success, they cannot model non-linear user-item relationships due to the linear nature of dot product. To overcome the limitation, neural network based recommenders were recently introduced [1, 9, 36, 38]. [9] combined a generalized matrix factorization component and a non-linear user-item interactions via a MLP architecture. [21, 24, 25] substituted the MLP architecture with the auto-encoder design. [35, 41] used memory augmentation to learn different user-item latent relationship. When non-existing users come with some observed interactions (i.e., recently created user accounts with some item interactions), the recommenders need to be rebuilt to generate their representations. To avoid these issues, current works encode users by combining the users’ consumed item embeddings in two main streams: (i) taking average of the consumed items’ latent representations [14, 17], or (ii) attentively summing [8] the consumed items’ embeddings.

Despite their success, general recommenders mostly consider all users’ unordered consumed items, and produce global/long-term user representations, which are supposed to be static, or changed slowly. Thus, they failed to capture the user’s dynamic behavior, that is captured by the user’s short-term preference (see Figure 1).

**Sequential Recommenders.** Sequential recommendation is known for its superiority to capture temporal dependencies between historical items [23]. Early works relied on Markov Chains to capture...
item-item sequential patterns [2, 5, 30]. Other works exploited the convolution architecture to capture more complex temporal dependencies [32]. These methods used short-term item dependencies to model a user’s dynamic interest. Other sequential recommenders focused on modeling long-term user preferences using RNN-based architectures [11, 20, 22, 40]. However, modeling either long-term user interests or short-term user interests is suboptimal since they concurrently affect a user’s intent (Figure 1). Recent works combined both long and short-term user preferences in real-valued representations to obtain satisfactory results [15, 42, 46]. Compared with the prior works which used real-valued representations, we propose Quaternion-based models to capture the user’s long-term and short-term interests. Quaternion was first introduced by [4] and it has recently shown its effectiveness over real-valued representations in NLP and computer vision domains [3, 27, 33, 47]. We acknowledge that QCF model [44] is the first Quaternion-based recommender. However, the authors used it as a simple extension of the matrix factorization method, where users and items are Quaternion embeddings. Thus, the benefits of Quaternion representation were not fully exploited in their network. Furthermore, they designed the model for a general recommendation problem, which has an inherent limitation of only modeling the user’s global interest.

3 PROBLEM DEFINITION

Denote \( U=\{u_1, u_2, ..., u_m\} \) as a set of all users where \( m = |U| \) is the total number of users, and \( P=\{p_1, p_2, ..., p_n\} \) as a set of all items where \( n = |P| \) is the total number of items. Bold versions of those variables, which we will introduce in the following sections, indicate their respective latent representations/embodiments. Each user \( u_i \in U \) consumes items in \( P \), denoted by a chronological list \( T^{(u_i)} \). We denote \( L^{(u_i)} \) as the chronological list of long-term consumed items of \( u_i \), and \( S^{(u_i)} \) as the chronological list of short-term consumed items of \( u_i \) (i.e., most recently consumed items in chronological order of the user \( u_i \)). \( L^{(u_i)} \cup S^{(u_i)} = T^{(u_i)} \). Note that bold versions of \( i, j, k \) are used to indicate the three imaginary parts of a Quaternion, while their subscript versions are used as indices.

In this work, we propose and build Quaternion-based recommender systems by using both long-term and short-term user interests, denoted as \( P(p_j|L^{(u_i)}, S^{(u_i)}) \). Under an assumption that \( L^{(u_i)} \) and \( S^{(u_i)} \) are independent given the target item \( p_j \), we model \( P(p_j|L^{(u_i)}, S^{(u_i)}) \) by modeling the user’s long-term interest \( P(p_j|L^{(u_i)}) \) and short-term interest \( P(p_j|S^{(u_i)}) \) separately by using two different Quaternion-based neural networks. Then, we automatically fuse the two models to build a more effective recommendation system.

4 PRELIMINARY ON QUATERNION

In this section, we cover important background on Quaternion Algebra and Quaternion Operators that we use to design our models.

Quaternion number: In mathematics, Quaternions are a hypercomplex number system. A Quaternion number \( X \) in a Quaternion space \( \mathbb{H} \) is formed by a real component \( r \) and three imaginary components as follows:

\[
X = r + ai + bj + ck, \tag{1}
\]

where \( ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j \). In Equa \( (1) \), \( r, a, b, c \) are real numbers \( \in \mathbb{R} \). Note that we can extend \( r, a, b, c \) to real-valued vectors to obtain a Quaternion embedding, which we use to represent each user/item’s latent features and conduct neural transformations. Operations on Quaternion embeddings are similar to Quaternion numbers.

Component-wise Quaternion Operators: Let \( f \) define an algebraic operator in real space \( \mathbb{R} \). The component-wise Quaternion operator \( f \) on two Quaternions \( X, Y \in \mathbb{H} \) is defined as:

\[
f(X, Y) = f(rx, ry) + f(ay, az) + f(bx, by) + f(cx, cz)k \tag{2}
\]

For instance, if \( f \) is an addition operator (i.e., \( f(a, b) = a + b \)), then \( f(X, Y) \) returns a component-wise Quaternion addition between \( X \) and \( Y \). If \( f \) is a dot product operator (i.e., \( f(a, b) = a^T b \)), then \( f(X, Y) \) returns a component-wise Quaternion dot product between \( X \) and \( Y \). A similar description is applied when \( f \) is either subtraction, scalar multiplication, product, softmax, or concatenate operator, etc.

Hamilton Product: The Hamilton product (denoted by the \( \otimes \) symbol) of two Quaternions \( X \in \mathbb{H} \) and \( Y \in \mathbb{H} \) is defined as:

\[
X \otimes Y = rxY - axY + byY - cxyY + (rY + axY + byY + cxyY) + (rxY + axY + byY + cxyY)k
\]

Activation function on Quaternions: Similar to [3, 28], we use a split activation function because of its stability and simplicity. Split activation function \( \beta \) on a Quaternion \( X \) is defined as:

\[
\beta(X) = \alpha(r) + \alpha(a)i + \alpha(b)j + \alpha(c)k \tag{4}
\]

where \( \alpha \) is any standard activation function for real values.

Concatenate four components of a Quaternion: concatenates all four Quaternion components into one real-valued vector:

\[
[X] = [rx, ax, bx, cx] \tag{5}
\]

5 OUR PROPOSED MODELS

Figure 3 shows an overview of our proposals. First, our Quaternion-based self-Attentive Long term user Encoding (QUALE) learns a user’s long-term interest by using long-term consumed items and the target item. Second, our Quaternion-based self-Attentive Short term user Encoding (QUASE) encodes the user’s short-term intent by using short-term consumed items and the target item. Then our Quaternion-based self-Attentive Long Short term user Encoding (QUALSE) fuses both of the user preferences by using a Quaternion-based gating layer. We describe each component as follows:

5.1 Quaternion-based self-Attentive Long Term user encoding (our QUALE model)

The most widely used technique for modeling the user long-term interests is the Asymmetric-SVD (ASVD) [17] model. Its basic idea is to encode each user and item by latent representations where the user representation is encoded by summing latent representations of the user’s interacted items. To an extent, we propose a Quaternion-based self-Attentive Long term user Encoding (QUALE). QUALE represents each user and each item as Quaternion embeddings. Then, we encode each user by attentively summing Quaternion embeddings of her interacted items as follows:
use Eq. (2) to multiply the results with the scaling factor $1/\sqrt{d}$ to eliminate the scaling effects. Then, we apply Component-wise Softmax (Eq. (2)) to obtain Quaternion attention scores as follows:

$$
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_l
\end{bmatrix} = \text{ComponentSoftmax} \left( \begin{bmatrix}
\frac{p_1}{\sqrt{d}} \\
\frac{p_2}{\sqrt{d}} \\
\vdots \\
\frac{p_l}{\sqrt{d}}
\end{bmatrix} \right)
$$

To obtain the attentive long-term user encoding $u_i^{(long)}$ of the user $u_i$, we first perform the component-wise product between the attention scores $[\alpha_1, \alpha_2, \ldots, \alpha_l]$ obtained in Eq. (7) with its corresponding item Quaternion embeddings $[p_1^{(long)}, p_2^{(long)}, \ldots, p_l^{(long)}]$. Then we sum them up to obtain $u_i^{(long)}$ as follows:

$$
u_i^{(long)} = \sum_{k=1}^{l} \alpha_k \times p_k^{(long)}$$

Our proposed Quaternion personalized self-attention mechanism vs. the existing self-attention mechanism: Our proposed Quaternion personalized self-attention mechanism is different from the self-attention mechanism that has been widely used in the NLP tasks in two aspects. First, unlike the prior work [45], which uses a single global context to assign attentive scores for different dialogue states, our attention mechanism provides personalized contexts for different users. In the recommendation domain, the long-term/general user interests are supposed to be changed slowly, but user interests are various across users. In other words, a user’s long-term context is quite static, but different from another user. Hence, using personalized contexts for different users is better than using a single global context, which is not personalized. Second, our attention mechanism adopts Hamilton product and works for Quaternion embeddings as input, instead of the real-valued embeddings like traditional self-attention mechanisms.

5.1.4 Output. We produce a long-term preference score $o_{ij}^{(long)}$ between the target user $u_i$ and the target item $p_j$ by computing the Component-wise dot product between the user long-term Quaternion encoding $u_i^{(long)}$ obtained in Eq. (8) and the target item Quaternion embedding $p_j^{(long)}$. This results in a Quaternion score.
To obtain a real-valued scalar preference score used in the parameter estimation phase, we compute the average of the scalar values of four Quaternion components by following [44]:

$$a_{ij}^{\text{(long)}} = \text{Average}(\text{ComponentDot}(u_i^{\text{(long)}}, p_j^{\text{(long)}}))$$

(9)

5.2 QUaternion-based self-Attentive Short term user Encoding (our QUASE model)

RNN-based models have gained a lot of attention because of their capability to capture item-to-item relationships [26, 40, 46]. However, due to its limitation in modeling a long sequence, we only exploit the RNN architecture to encode a user’s short-term interest. Recently, [28] has introduced a Quaternion LSTM (QLSTM) model and has shown its efficiency and effectiveness over a traditional real-valued LSTM model. However, QLSTM used only the last hidden state as a latent summary of the input, which is suboptimal. To an extent, we propose a Quaternion-based self-Attentive LSTM model to learn a user’s short-term interest. We name our proposal as a QUaternion-based self-Attentive Short term user Encoding (QUASE). QUASE has 4 layers: Input, Quaternion Embedding, Encoding, and Output layers. We describe each layer as follows:

5.2.1 Input. A target item $p_j$, and the chronological list of $s$ short-term consumed items $S^u(t)$ of the target user $u_t$ with $|S^u(t)| = s$, where $s$ represents the maximum number of short-term items among all the users in a dataset. If a user has consumed less than $s$ items, we pad the list with zeroes until its length reaches $s$.

5.2.2 Quaternion Embedding layer. It holds an item Quaternion embedding matrix $p_j^{\text{(short)}} \in \mathbb{R}^{d \times d}$. By passing the target item $p_j$, and $s$ short-term items in the $S(t)$ of the target user $u_t$ through $p_j^{\text{(short)}}$, we obtain their corresponding Quaternion embeddings $p_j^{\text{(short)}}$, and $p_1^{\text{(short)}}$, $p_2^{\text{(short)}}$, ..., $p_s^{\text{(short)}}$.

5.2.3 Encoding layer. In this layer, we adapt the recently introduced Quaternion-based LSTM to model the item-item sequential transition. Denote $p_j^{\text{(short)}}$ is the Quaternion embedding of the $t$th short-term item $p_j \in S(t)$ ($t = 1, s$). Let $f_t$, $i_t$, $o_t$, and $c_t$ be the forget gate, input gate, output gate, and cell state, respectively, and the hidden state of a Quaternion LSTM cell at time step $t$, respectively. We compute these variables as follows:

$$f_t = \sigma(W_f \otimes p_j^{\text{short}} + R_f \otimes h_{t-1} + g_f)$$

$$i_t = \sigma(W_i \otimes p_j^{\text{short}} + R_i \otimes h_{t-1} + g_i)$$

$$o_t = \sigma(W_o \otimes p_j^{\text{short}} + R_o \otimes h_{t-1} + g_o)$$

$$c_t = f_t \times c_{t-1} + i_t \times \text{tanh}(W_c \otimes p_j^{\text{short}} + R_c \otimes h_{t-1} + g_c)$$

$$h_t = o_t \times \text{tanh}(c_t)$$

(10)

where $W_f$, $R_f$, $W_i$, $R_i$, $W_o$, $R_o$, $W_c$, $R_c$, $g_f$, $g_i$, $g_o$, $g_c$ are Quaternion weight matrices, $g_f$, $g_i$, $g_o$, $g_c$, $h_t$ are Quaternion bias vectors. $f_t$, $i_t$, $o_t$, $c_t$, $h_t$ are Quaternion vectors. The “$\times$” sign denotes a component-wise product operator, which is calculated using Equa (2). sigmoid $\sigma$ and tanh are split activation functions and are computed using the Equa (4).

Using Equa (10), given $s$ short-term consumed items $p_1, p_2, ..., p_s$, we obtain their respective output Quaternion hidden states $h_1, h_2, ..., h_s$. Then, we propose a Quaternion self-attention mechanism to combine all $s$ output Quaternion hidden states before using it to predict the next item. Different from the long-term user preferences where they are supposed to be static or changed very slowly, the short-term user interests are dynamic and changed quickly. Hence, using a static user context for each user to make personalized attention like what we did for the QUASE model is not ideal. Instead, we define a Quaternion global context vector to capture the sequential transition patterns from item to item among all the users. Denote $q$ as a Quaternion global context vector, the Quaternion-based self-attention score of each hidden state $h_t$ is measured by:

$$a_{ij}^{\text{(short)}} = \text{ComponentSoftmax}(\left[ h_1 \otimes q / \sqrt{d}, h_2 \otimes q / \sqrt{d}, ..., h_s \otimes q / \sqrt{d} \right])$$

(11)

where $a_1^{\text{(short)}}$, $a_2^{\text{(short)}}, ..., a_s^{\text{(short)}}$ are Quaternion numbers. To achieve the final short-term user Quaternion encoding, we perform a component-wise product between the Quaternion hidden states and their respective Quaternion attention scores, followed by a Hamilton product with a Quaternion weight matrix $W$ and the split activation function tanh:

$$u_i^{\text{(short)}} = \text{tanh}(W \otimes \left[ \sum_{t=1}^{s} a_i^{\text{(short)}} \times h_t \right])$$

(12)

Note that we also designed a Quaternion self-Attentive GRU, but its performance was slightly worse than the Quaternion self-Attentive LSTM (see Table 2 in Section 6). Thus, we only described the Quaternion self-Attentive LSTM due to space limitation.

5.2.4 Output. Similar to Equa (9), we produce the user $u_t$ short-term preference score $a_{ij}^{\text{(short)}}$ over the target item $p_j$ as follows:

$$a_{ij}^{\text{(short)}} = \text{Average}(\text{ComponentDot}(u_i^{\text{(short)}}, p_j^{\text{(short)}}))$$

(13)

5.3 QUaternion-based self-Attentive Long Short term user Encoding (QUALSE): a Fusion of QUASE and QUALE models

In this part, we aim to combine both user’s long-term and short-term preferences modeling parts into one model, namely QUALSE, fusing QUASE and QUALE models. Inspired by the gated mechanism in LSTM [12] to balance the contribution of the current input and the previous hidden state, we propose a personalized Quaternion gated mechanism to fuse the long-term and short-term user interests learned in QUASE and QUALSE models. Our personalized gating proposal is different to the traditional gating mechanism in two folds. First, gating weights in our proposal are in Quaternion space and the transformations are computed using the Hamilton product. Second, as users’ behaviors differ from a user to another user, we additionally input the target user embeddings $u_t$ to let the gating layer assign personalized scores for different users. The long-term and short-term interest fusion is computed as follows:

$$y_{ij}^{\text{(long)}} = \sigma(W_o^{(1)} \otimes (u_i^{\text{(long)}}, u_j^{\text{(long)}}, W_s^{(2)} \otimes u_i + W_s^{(3)} \otimes p_j)$$

$$o_{ij} = W_o^{(2)} \left[ (1 - y_{ij}^{\text{(long)})} \times (u_i^{\text{(short)}}, p_j^{\text{(short.ListView}}) \right]$$

(14)
\(D\) and \(j\) are Quaternion vectors respectively. To compute the long-term gate \(\lambda_{ij}^{(long)}\), \(u_i^{(long)}\) and \(p_j^{(long)}\) are introduced as an additional user context Quaternion embedding and a target item context Quaternion embedding to let the model know which long-term or short-term interests are more relevant. To measure the final output \(o_{ij}\), since \(\lambda_{ij}^{(long)}\) is a Quaternion vector while \(\lambda_{ij}^{(long)}\) and \(\lambda_{ij}^{(short)}\) are scalar values, we reconstruct the user’s long-term interest by computing \(u_i^{(long)} \times p_j^{(long)}\) and the short-term interest by measuring \(u_i^{(short)} \times p_j^{(short)}\), which are also Quaternion vectors. Finally, to combine multiple dimensional features from the weighted long-term and short-term interest Quaternion vectors, we concatenate all their components, denoted by \([\cdot\,\cdot\,\cdot]\) (Equa (5)), and use two real-valued weight vectors \(W_o^{(1)}\) and \(W_o^{(2)}\) to produce a fused preference score as a scalar real number. Note that in QUALE, QUASE and QUALE hold separated item memory to increase the their flexibility.

5.4 Parameter Estimation

5.4.1 Training with Bayesian Personalized Ranking (BPR) loss. Given a Quaternion matrix \(E \in \mathbb{R}^{(m+n) \times d}\) as the Quaternion embeddings of all users and items in the system, and \(\Theta\) as other parameters of the model, we aim to minimize the following BPR loss function:

\[
L_{BPR}(D|E, \Theta) = \arg\min_{E, \Theta} \left( - \sum_{(i, j^+, j^-)} \log(o_{ij^+} - o_{ij^-}) + \lambda_{\Theta} \|E\|_2 + \lambda_{E} \|E\|_2 \right)
\]

where \(\lambda_{\Theta}\) and \(\lambda_{E}\) are regularization hyper-parameters.

5.4.2 Training with Quaternion Adversarial attacks. Previous works have shown that neural networks are vulnerable to adversarial noise [7, 19]. Therefore, to increase the robustness of our models, we propose a Quaternion Adversarial attack on BPR loss, namely QABPR. QABPR inherits from traditional adversarial attacks for computer visions [19] and recommendation systems [7] but differs from them: QABPR applies for Quaternion space, while the formers apply for real-valued space. To our best of knowledge, ours is the first work using adversarial training on Quaternion space in the recommendation domain.

In QABPR, we first define learnable Quaternion perturbation noise \(\delta\) on user and item Quaternion embeddings. Then, we perform the Quaternion component-wise addition (Equa (2)) to obtain crafted Quaternion embeddings. The learnable Quaternion noise \(\delta\) is optimized such that the model mis-ranks between positive items and negative items (i.e. negative items have higher preference scores than positive items). Particularly, a max player learns \(\delta\) by maximizing the following cost function under the \(L_2\) attack:

\[
\delta = \epsilon \frac{\nabla_{\delta} L_{adv}(D|E^* + \delta, \Theta^*)}{\|\nabla_{\delta} L_{adv}(D|E^* + \delta, \Theta^*)\|_2}
\]

where \(\epsilon\) is a noise magnitude hyper-parameter. \(E^*\) and \(\Theta^*\) are optimal values of \(E\) and \(\Theta\) that are pre-learned in Equa (15) and are fixed in Equa (16). \(E^* + \delta\) is the crafted Quaternion embeddings. \(\lambda_{\Theta}\) and \(\lambda_{E}\) in Equa (15) are ignored in Equa (16) as they become constant terms. \(\lambda_{\delta}\) is the noise regularization term.

Solving Equa (16) is expensive. Hence, we adopt the Fast Gradient Method [19] to approximate \(\delta\) as follows:

\[
\delta = \epsilon \nabla_{\delta} L_{adv}(D|E^* + \delta, \Theta^*)
\]

Then, a min player aims to minimize the following cost functions that incorporate both non-adversarial and adversarial examples:

\[
L_{QABPR}(D|E, E^* + \delta, \Theta, \lambda)
\]

\[
= \arg\min_{E, \Theta} \left( L_{BPR}(D|E, \Theta) + \lambda_{adv} \|E^* + \delta, \Theta\|_2 \right)
\]

where \(\delta\) is the adversarial noise that is already learned in Equa (17), and is fixed in Equa (18). \(\lambda_{adv}\) is a hyper-parameter to balance the effect of the partial adversarial loss. Training QABPR now becomes playing a minimax game, where the mini and maxi players play alternately. We stop the game after a fixed number of epochs (i.e. 30 epochs) and report results based on the best validation performance.

Note that we name our QUALE, QUASE and QUASE trained with QABPR loss as AQUALE, AQUASE, and AQUASE with “A” denotes “adversarial”, respectively.

### Table 1: Datasets’ statistics with \# of long-term items \(l\).

| Dataset          | \# of users | \# of items | \# of actions (density %) | \# of long-term items |
|------------------|-------------|-------------|---------------------------|-----------------------|
| Toys Games       | 36k         | 55k         | 251k (0.013%)             | 1112                  |
| Cellphone Accessi| 47k         | 45k         | 262k (0.012%)             | 109                   |
| Pet Supplies     | 25k         | 23k         | 160k (0.027%)             | 176                   |
| Video Games      | 25k         | 20k         | 196k (0.040%)             | 856                   |
| Apps for Android | 79k         | 18k         | 555k (0.038%)             | 478                   |
| Yelp             | 22k         | 21k         | 481k (0.104%)             | 930                   |

6. EMPIRICAL STUDY

In this section, we design experiments to answer the following research questions:

- **RQ1**: How do our proposals work compared to the baselines?
- **RQ2**: How do a user’s long-term, short-term preference encoding models and the fused model perform?
- **RQ3**: Is using Quaternion representation helpful and why?
- **RQ4**: Are the gating fusion mechanism and the Quaternion BPR adversarial training helpful?

6.1 Datasets

We evaluate all models on six public benchmark datasets collected from two real-world systems as follows:

- **Amazon datasets** [6]: As toplevel product categories on Amazon are treated as independent datasets [15], we use 5 different Amazon category datasets to vary the sparsity, variability,
We compared our proposed models with \textit{11} strong state-of-the-art recommendation models as follows:

- **AASVD**: It is an attentive version of the well-known Asymmetric SVD model (ASVD) [17], where real-valued self-attention is applied to measure attentive contribution of previously consumed items by a user.

- **QCF** [44]: It is a state-of-the-art recommender that represents users/items by Quaternion embeddings.

- **NeuMF++** [9]: It models non-linear user-item interactions by using a MLP and a Generalized MF (GMF) component. We pretrained MLP and GMF to obtain NeuMF’s best performance.

- **NAIS** [8]: It is an extension of ASVD where contribution of consumed items to the target item is attentively assigned. We adopt \textit{NAIS}_{prod} version as it led to its best results.

- **FPMC** [30]: It is a state-of-the-art sequential recommender. It uses the first-order Markov to model the transition between the next item and the previously consumed items.

- **AGRU** [8]: It is an extension of the well-known GRU4Rec [11], where we use an attention mechanism to combine different hidden states. We experiment with two attention mechanisms: real-valued self-attention, and real-valued \textit{prod attention} proposed by [8]. Then we report its best performance.

- **ALSTM**: It is a LSTM based model. Similar to AGRU, we experiment with the real-valued self-attention and the \textit{prod attention} [8], and then report its best results.

- **Caser** [32]: It embedded a sequence of recently consumed items into an “image” in time and latent spaces, and uses convolution neural network to learn sequential patterns as local features of an image using different horizontal and vertical filters.

- **SASRec** [15]: It is a strong sequential recommender model. It uses the self-attention mechanism with a multi-head design to identify relevant items for next predictive items.

- **SLi-Rec** [42]: It uses a time-aware controller to control the state transition. Then it uses an attention-based framework to fuse a user’s long-term and short-term preferences.

- **ALSTM+AASVD**: It is our implementation that resembles the Sli-Rec’s same architecture as our proposed Quaternion fusion approach, except that it uses Euclidean space instead of Quaternion space. The purpose of implementing and using it as a baseline is to present the effectiveness of our framework and Quaternion representations over the real-valued representations.

First four baselines (AASVD, QCF, NeuMF++, and NAIS) are classified as user’s long-term interest encoding models. Next four baselines (FPMC, AGRU, ALSTM, and Caser) are user’s short-term interest encoding models, and SASRec, SLi-Rec, and ALSTM+AASVD encode both user’s long-term and short-term intents. Note that we performed an experiment with \textit{Dien} [46] (i.e. a long short-term
modeling baseline) based on the authors’ public source code, which produced surprisingly low results, so we omit its detailed results. We also experimented with ASVD, LSTM, GRU and Quaternion LSTM but do not report their results due to space limitation and their worse results. Similarly, we omit BPR [29] and FISM [14] results due to their less impressive performance.

6.3 Experimental Settings

Protocol: We adopt a well-known and practical 70/10/20 splitting proportions to divide each dataset into train/validation (or development)/test sets [21, 34]. All user-item interactions are sorted in ascending order in terms of the interaction time. Then, the first 70% of all interactions are used for training, the next 10% of all interactions are used for development, and the rest is used for testing. We follow [39, 41] to sample 1,000 unobserved items that the target user has not interacted before, and rank all her positive items with these 1,000 unobserved items for testing models.

Evaluation metrics: We evaluate the performances of all models by using two well-known metrics: Hit Ratio (HIT@N), and Normalized Discounted Cumulative Gain (NDCG@N). HIT@N measures whether all the test items are in the recommended list or not, while NDCG@N takes into account the position of the test items, and assigns higher scores if test items are at top-rank positions.

Hyper-parameters Settings: All models are trained with Adam optimizer [16]. A learning rate is chosen from \([0.001, 0.0005]\), and regularization hyper-parameters are chosen from \([0, 0.1, 0.001, 0.0001]\). An embedding size \(d\) is chosen from \([32, 48, 64, 96, 128]\). Note that for Quaternion embeddings, each component value is a vector of size \(d^2/4\). The number of epochs is 30. The batch size is 256. The number of negative samples per one positive instance is 4 for training models. The settings of Caser, NAIS, SASRec are followed by their reported default settings. In training with QABPR loss, the regularization \(\lambda_{adv}\) is set to 1. The noise magnitude \(\epsilon\) is chosen from \([0.5, 1, 2]\). The adversarial noise is added only in training process, and is initialized as zero. All hyper-parameters are tuned by using the validation set.

6.4 Experimental Results

6.4.1 RQ1: Performance comparison. Table 2 shows that our proposed fused models QUALSE and AQUALSE outperformed all the compared baselines. On average, QUALSE improved Hit@100 by 5.65% and NDCG@100 by 8.44% compared to the best baseline’s performances. AQUALSE gains additional improvement over QUALSE, enhancing Hit@100 by 7.39% and NDCG@100 by 10.83% on average compared to the best baseline. The improvement of our proposals over the baselines is significant under the Directional Wilcoxon signed-rank test (\(p\)-value < 0.015). We also observed similar results on all six datasets when we measure Hit@1 and NDCG@1. In particular, our QUALSE improved Hit@1 by 6.87% and NDCG@1 by 8.71% on average compared with the best baseline. AQUALSE improved Hit@1 by 8.43% and NDCG@1 by 10.27% on average compared with the best baseline, confirming its consistent effectiveness.

6.4.2 Varying top-N recommendation list and embedding size: To further provide detailed effectiveness of our proposals, we compare QUALSE and AQUALSE models with the top-5 baselines when varying the embedding size from \([32, 48, 64, 96, 128]\) and the top-N recommendation list from \([10, 20, 50, 100]\).
is due to the fact that SASRec models user’s long-term and short-term interests implicitly and concurrently by using the Transformer multi-head attention mechanism. But, SLi-Rec, ALSTM+AASVD, and our proposals model the two preferences explicitly and separately, and then combine them later on, increasing flexibility. Note that, although SLi-Rec employed a time-aware attentional LSTM to better model the user’s short-term preferences, our ALSTM+AASVD implementation works slightly better than SLi-Rec due to its two distinct properties: (i) the personalized self-attention in AASVD, where each user is parameterized by her own context vector, and (ii) the personalized gating fusion.

6.4.4 **Q3: Is using Quaternion representation helpful?** In Table 2, we compare different model pairs: AASVD vs. QUALE, ALSTM vs. QUASE (LSTM), AGRU vs. QUASE (GRU), and ALSTM+AASVD vs. QUALSE. Two methods under the same pair have similar architecture (again, ALSTM+AASVD was implemented by us, following our QUALSE architecture to show effectiveness of Quaternion representation). But, the first method of each pair uses real-valued representations and the second method of each pair uses Quaternion representations. Table 2 shows that QUALE works better than AASVD. In six datasets, on average, QUALE improves Hit@100 by 5.60% and NDCG@100 by 7.11% compared to AASVD. Similarly, we observe the same patterns from the other three model pairs. Moreover, when comparing our long-term encoding QUALE and QUALSE models with other long-term encoding baselines (a)-(e), our models outperformed the baselines, improving Hit@100 by 5.16% and 6.55%, and enhancing NDCG@100 by 7.11% and 9.83%, respectively. Similarly, our short-term encoding QUASE and AQUASE using LSTM also work better than other short-term encoding baselines (f)-(h), improving HIT@100 by 4.75% and 8.09%, and enhancing NDCG@100 by 10.57% and 15.33%, respectively. All of these results confirm the effectiveness of modeling user’s interests by using Quaternion representations over Euclidean representations.

**Why Quaternion representations help improve the performance?** Since attention mechanism is the key success in deep neural networks [37], we analyze how our models assign attention weights compared to their respective real-valued models. We first measure the item-item Pointwise Mutual Information (PMI) scores (i.e. $PMI(j, t) = \log \frac{P(j|t) \cdot P(t|j)}{P(j) \cdot P(t)}$) using the training set. The PMI score between two items $(j, t)$ gives us the co-occurrence information between item $j$ and item $t$, or how likely the target item $j$ will be preferred by the target user when the item $t$ is already in her consumed item list. We perform softmax on all item-item PMI scores. Then, we compare with the generated attention scores from our proposed models and ones from their respective real-valued baseline models.

Figure 5: Density distribution of item-item similarity scores in train/val/test sets of Pet Supplies and Yelp datasets.

![Figure 5](image)

**6.4.5 Q4: Effect of the personalized gated fusion and the QABPR loss?** Table 2 shows that in real-valued representations, ALSTM+AASVD works better than AASVD and ALSTM in all six datasets. Similarly, in Quaternion representations, the fused QUALSE model generally works better than its two degenerated QUALE and QUASE models. In the six datasets, both QUASE and AQUASE perform better than their degenerated (adversarial) versions, improving 2% on average w.r.t both HIT@100 and NDCG@100. The results confirm the effectiveness of fusing long-term and short-term user preferences in both of QUASE and AQUASE.

We further compare our gating fusion with a weight fixing method, where we vary a contribution score $c \in [0, 1]$ for the user’s short-term preference encoding part and $1-c$ for the long-term part. We see that the gating fusion improves 4.82% on average over six datasets compared to the weight fixing method, again confirming the effectiveness of our personalized gating fusion method.

**Is Quaternion Adversarial training on BPR loss helpful?** We compare our proposed models training with BPR loss (i.e. QUALE, QUASE (LSTM), and QUASE models) and our proposed models training with QABPR loss (i.e. AQUALE, AQUASE (LSTM), and AQUASE)
First, we observe that AQUASE boosted QUASE performance by a large margin: improving HIT@100 by 3.2% and NDCG@100 by 4.29% on average in the six datasets. AQUALE and AQUALSE also improve QUALE and QUALSE by 1.92% and 1.91% on average of both HIT@100 and NDCG@100 over six datasets, respectively. These results show the effectiveness of the adversarial attack on Quaternion representations with our QABPR loss.

7 CONCLUSION

In this paper, we have shown that user’s short-term and long-term interests are complementary and both of them are indispensable. We fully utilized Quaternion space and proposed three novel Quaternion-based models: (1) a QUALE model learned the user’s long-term intents, (2) a QUASE model learned the user’s short-term interests, and (3) a QUALSE model fused QUALE and QUASE to learn both user’s long-term and short-term preferences. We also proposed a Quaternion-based Adversarial attack on Bayesian Personalized Ranking (QABPR) loss to improve the robustness of our proposals. Through extensive experiments on six real-world datasets, we showed that our QUALSE improved 6.87% at HIT@1 and 8.71% at NDCG@1, and AQUALSE improved 8.43% at HIT@1 and 10.27% at NDCG@1 on average compared with the best baseline. Our proposed models consistently achieved the best results when varying top-N (e.g., HIT@100 and NDCG@100). These results show the effectiveness of our proposed framework.

8 ACKNOWLEDGMENTS

This work was supported in part by NSF grant CNS-1755536, AWS Cloud Credits for Research, and Google Cloud.

REFERENCES

[1] Travis Ebesu, Bin Shen, and Yi Fang. 2018. Collaborative memory network for recommendation systems. In SIGIR. 515–524.
[2] Shanshan Feng, Xutao Li, Yifeng Zeng, Gao Cong, Yeow Meng Chee, and Quan Yuan. 2015. Personalized ranking metric embedding for next new POI recommendation. In IJCAI. 2069–2075.
[3] Chase J Gaudet and Anthony S Maida. 2018. Deep quaternion networks. In IJCNN. 1–8.
[4] William Rowan Hamilton. 1844. LXXVIII. On quaternions; or on a new system of imaginaries in algebra. To the editors of the Philosophical Magazine and Journal of the London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 25, 169 (1844). 489–495.
[5] Ruining He, Wang-Cheng Kang, and Julian McAuley. 2017. Translation-based recommendation. In RecSys. 161–169.
[6] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In WWW. 507–517.
[7] Xiangnan He, Zhankui Liu, Xiaoou Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In SIGIR. 355–364.
[8] Xiangnan He, Zhankui Liu, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. 2018. NAIS: Neural attention item similarity model for recommendation. IEEE TKDE 30, 12 (2018), 2354–2366.
[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqi Nie, and Tat-Seng Chua. 2017. Neural collaborative filtering. In WWW. 173–182.
[10] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In SIGIR. 549–558.
[11] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domenkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In ICLR.
[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.
[13] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In ICDM. 263–272.
[14] Santosh Kabbur, Xia Ning, and George Karypis. 2015. Fism: factored item similarity models for top-n recommender systems. In KDD. 659–667.
[15] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In ICDM. 197–206.