

Attention Pooling for Semantic Sessions and Neural-Gate Neuron Routing

Xiaonan Wang

Search-Ads Algorithms Related Industry Project When
Working as a Machine Learning Scientist
(During Period: 2020.09~2022.07)

Ranking Model Structure

DNN

Output A

Output B

Tower A

Tower B

Reference:

[1] Ma, Jiaqi, et al. "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.

Mixture of
Expert 0 and 1

Concat

Mixture of
Expert 1 and 2

Embedding

Slot

Expert 0

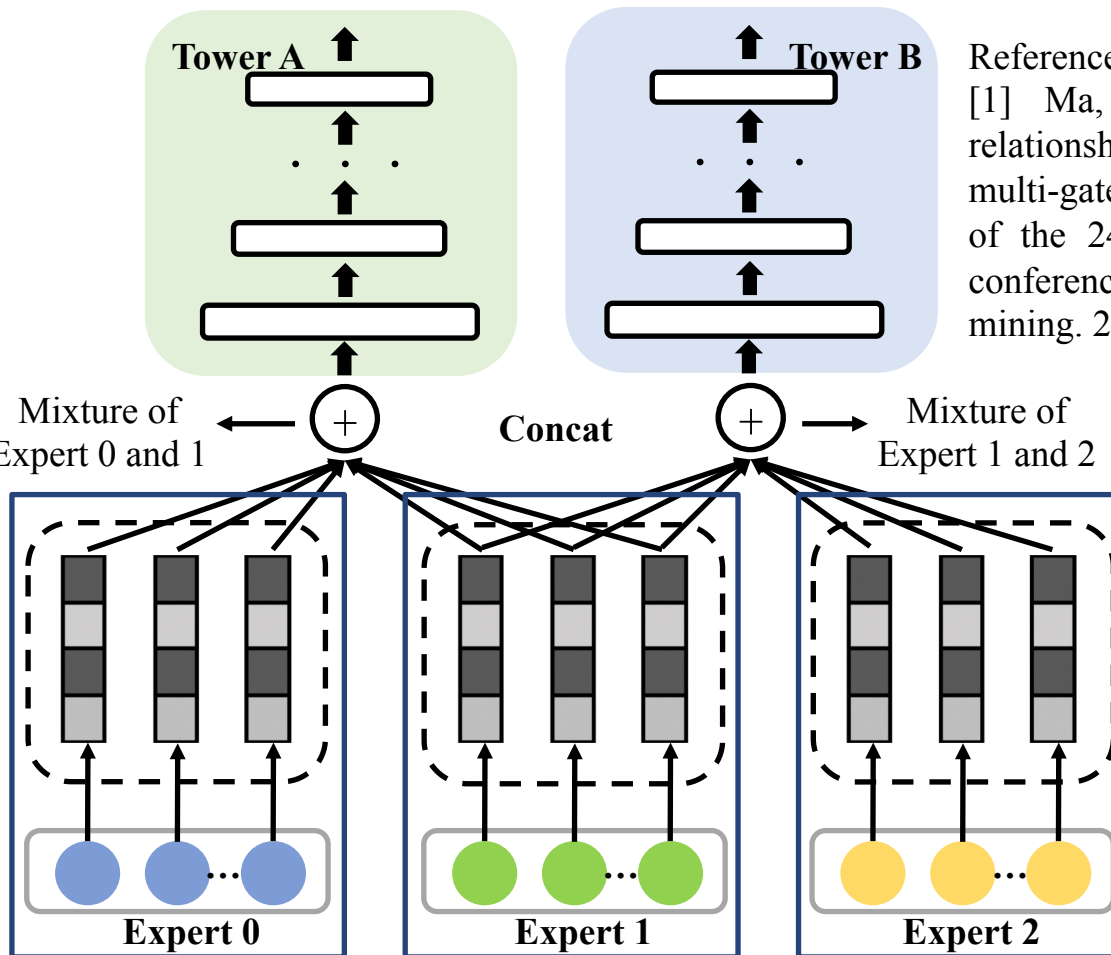
Expert 1

Expert 2

Memory features for
target A:
memory id, ip/region,
cross product feature

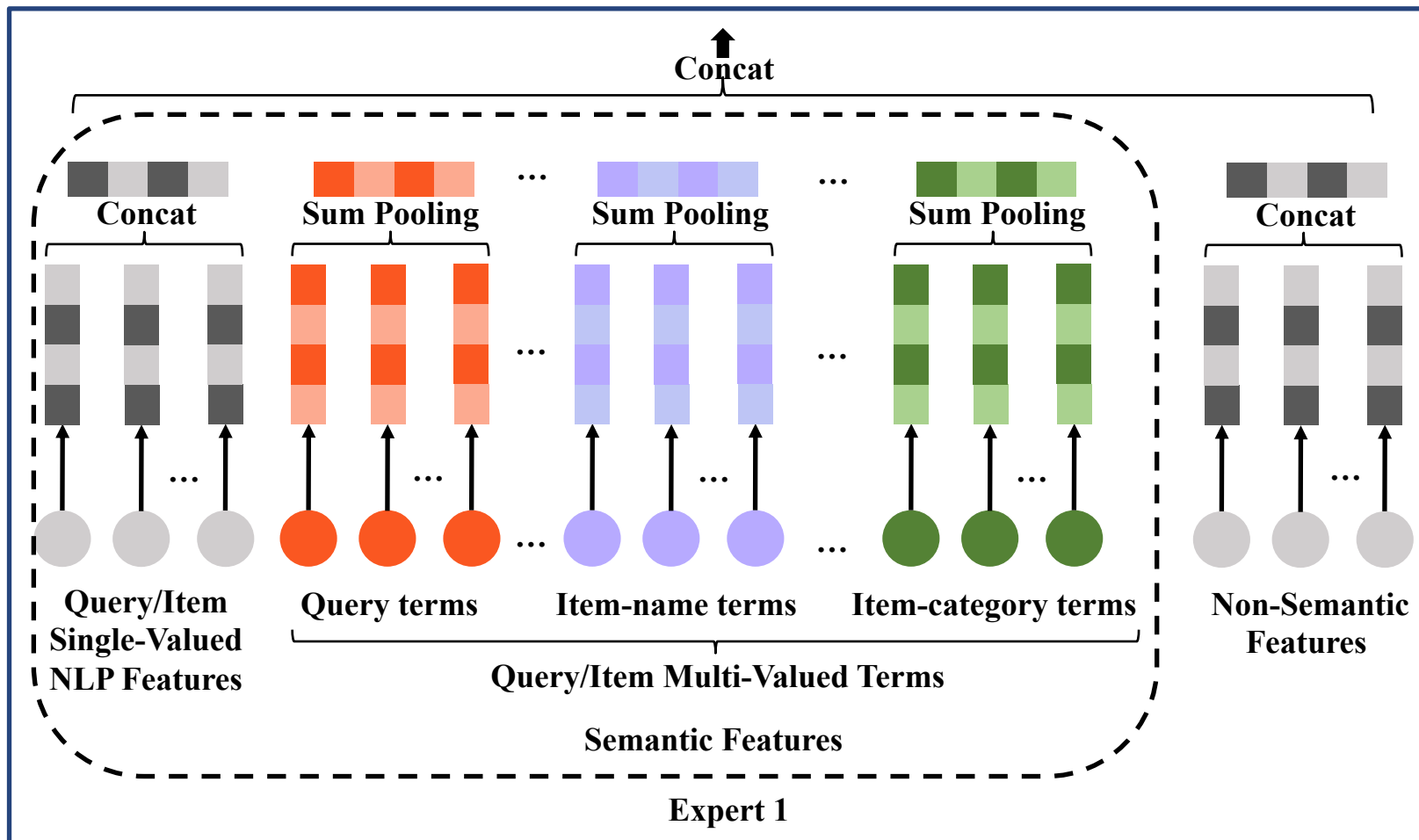
Generalized features
shared by target A and B:
semantics, category, user
session, portrait, etc.

Memory features for
target B:
memory id, ip/region,
cross product feature



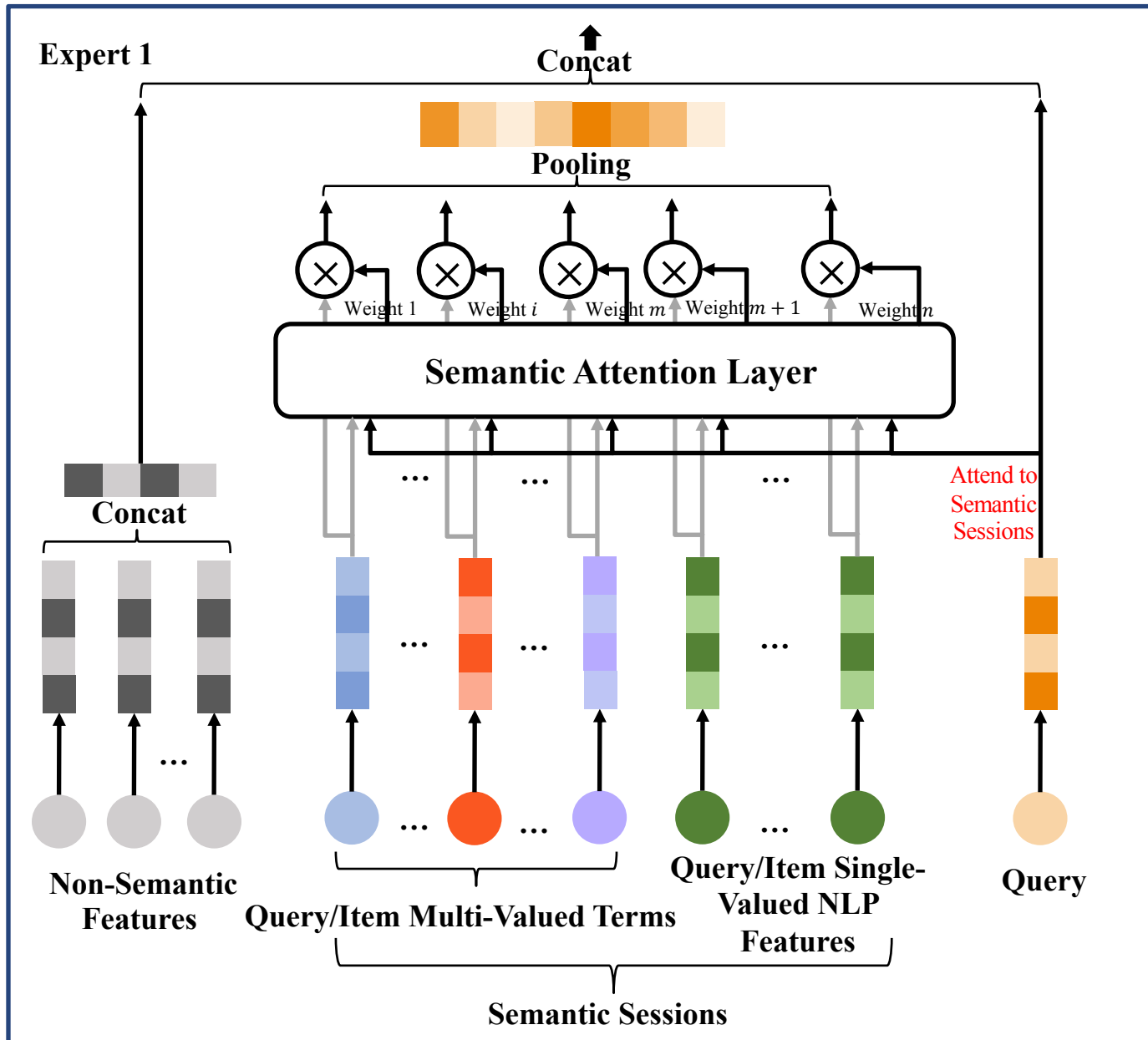
Semantic Attention Pooling

Base



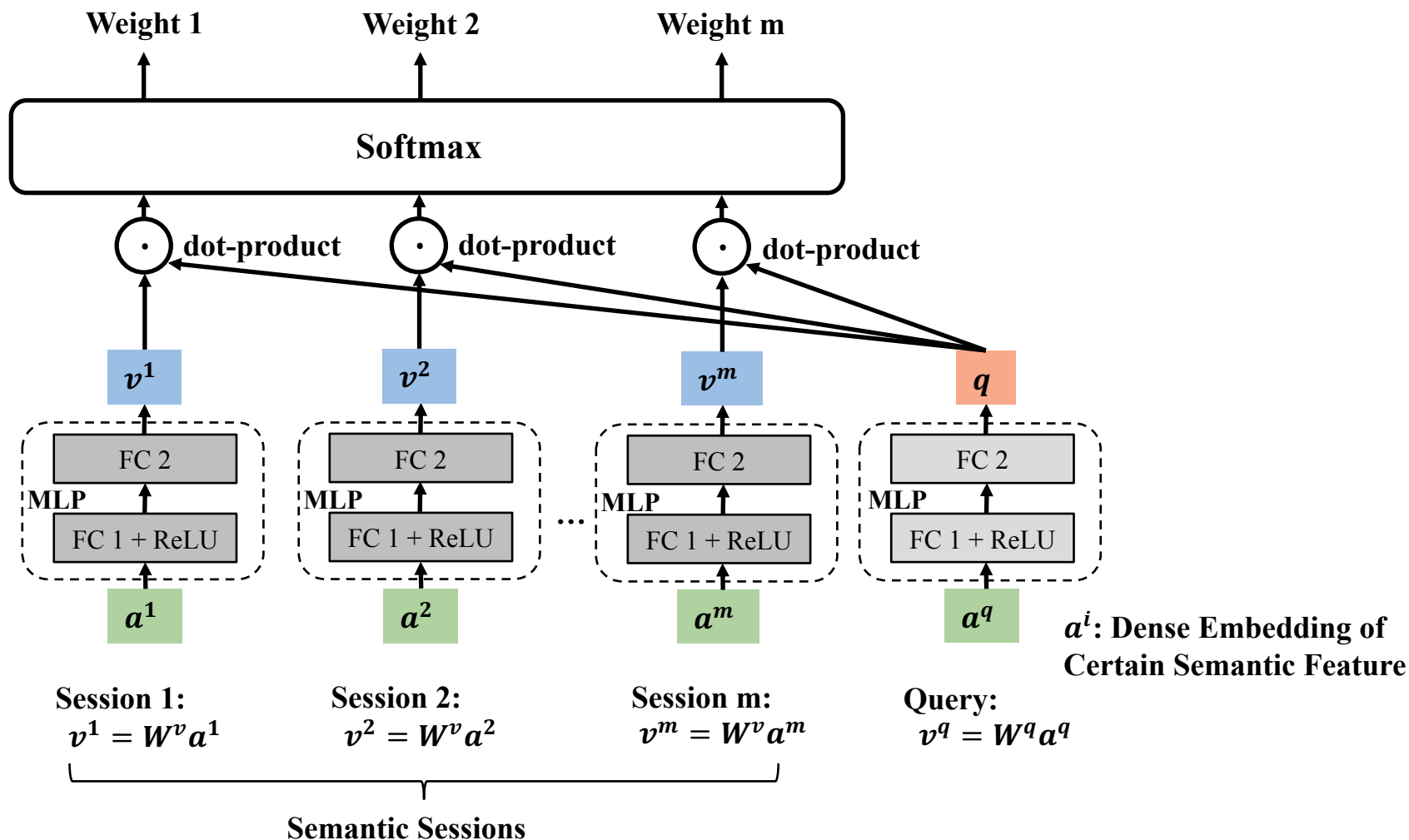
- Query/Item Multi-Valued Terms include: Query terms, Item-name terms, Item-Description Terms, Item-First-Category Terms, Item-Second-Category Terms, Item-Third-Category Terms.
- Query/Item Single-Valued Features: Single-Valued NLP Features, such as Query (NER), Item NER, etc.

Semantic Attention Pooling

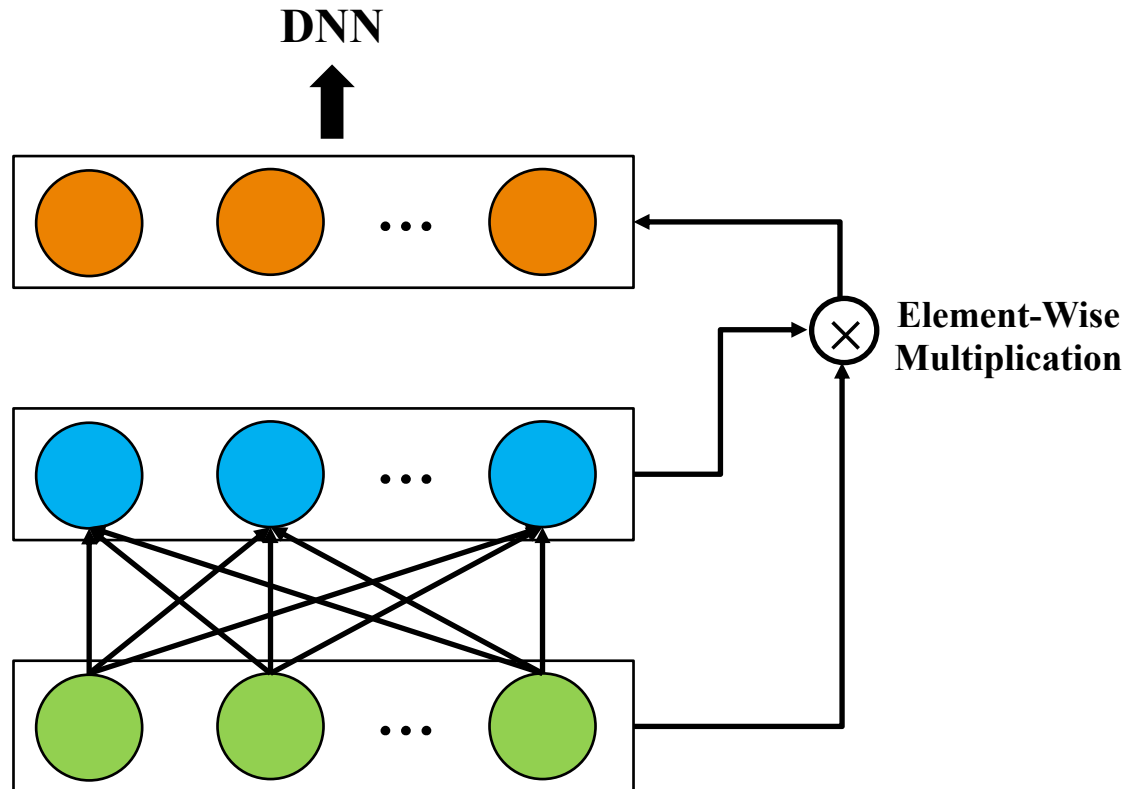


Semantic Attention Pooling

Semantic Attention Layer



Neural-Gate Neuron Routing



Summarization

Attention Pooling for Semantic Sessions

- Attention Weighted Sum Pooling for Semantic Sessions
- Query attend to semantic sessions to get relevance between query and each session. The relevance, which means the similarity to user's intention, will be used as sessions' weights.
- By doing this, we had achieved the pruning of the model (decreased from 30 semantic embedding-vectors to 2 semantic embedding-vectors as dnn's inputs) to further avoid overfitting.

Neural-Gate Neuron Routing

- “Environmental awareness” of embedding vector granularity.