Bonprix

Interview - Data Scientist

Sentence to Session Embeddings

HELLO!

I am Shrikanth Singh

Native Indian, 27 years old

Graduate in Computer Science from Uni Passau

Data Scientist @ Conrad Electronics

Literature Survey

References

1. Pattern2Vec: Representation of clickstream data sequences for learning user navigational behavior
2. Event2vec: Learning Representations of Events on Temporal Sequences

Inference

- Users’ behaviour is analogous to real kitchen funnel (ToFU - Site visit and BoFU - Purchase)
- Ideas inherited from NLP domain to model the user interactions on the web platform
  - Sentence = Session
  - Words in Sentence = Events in Session
  - Weighted averaging of event embeddings in a session
  - Pattern2Vec and Event2vec utilize concepts from Word2Vec and Node2Vec
- Incorporation of temporal order, event significance using ranking, relationship between events.

Session encoding using Sentence Transformer

Background

- SBERT modification of the BERT network using siamese and triplet networks that is able to derive semantically meaningful sentence embeddings.
- Faster inference computation compared to BERT, RoBERTa and other SOTA architectures.
- Outperforms Avg. BERT embeddings, Glove and Universal sentence encoder

Advantages

- Alleviates the need of explicitly modelling temporal and contextual dependencies
- Self attention mechanism - Position invariant i.e shuffled events does not impact the network
- Positional embedding - Encodes the information of the position of the event in the session
- Understands long range dependencies between events
- Multiple use cases served from same output embeddings - Semantic search, Anomaly detection, Customer intention and Churn prediction

Not just fine-tuning - Target domain (Click stream data) is incompatible with the dataset used for pre-training BERT

Regression function using Siamese BERT-Network Architecture

Objective

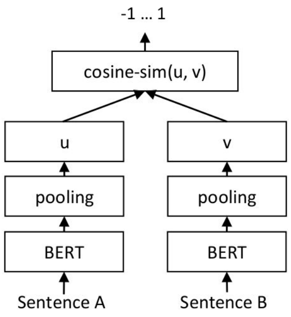
The cosine similarity between the two session embeddings u and v is computed and mean squared loss is used as the objective function.

Model and Dataset format

- Model - bert-base-uncased, Maximum sequence length - 512 events per session
- Data - Collection of pair of sessions (sentences) from different users
- Label - Correlation between the pair of sessions (Need not necessarily be Cosine similarity)

Output

- Session encoding using trained model - N x d matrix where N is number of sessions and d is embedding dimension i.e 768



Unsupervised Sentence Embedding Learning - TSDAE

Objective

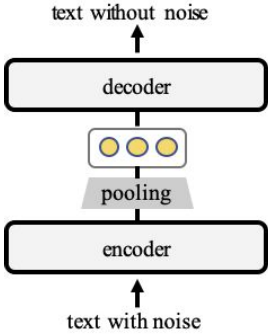
- TSDAE introduces noise to input sessions by deleting or swapping events. These damaged sessions are encoded by the transformer model into session vectors.
- Another decoder network then attempts to reconstruct the original input from the damaged session encoding.

Model and Dataset format

- Model - bert-base-uncased, Maximum sequence length - 512 events per session
- Data - Collection of sessions (sentences) from different users

Output

- Session encoding using trained model - N x d matrix where N is number of sessions and d is embedding dimension i.e 768



Concept Evaluation

Efforts and Success prospects

Certain processes need to be implemented irrespective of the approach used for encoding the user sessions.

- Data cleaning
  - Removal of noisy sessions using CDF of Maximum time gap between two arbitrary events
  - Eliminate high time intervals and long event sequence
- Creating cleansed sequence data - If number of events > Choice of maximum session length for model
  - Create sub-sessions by splitting based on Information theory to preserve the context of behaviour
  - Dropping redundant consecutive events
  - Allow limited looping of events in the same session
- Labeling - In case of supervised training approach additional efforts will be required to determine the correlation between the session. Not required in case of TSDAE.
- Unlike Pattern2Vec, attention model do not require additional engineering for temporal and semantic information. Therefore the proposed transformer approach should at least produce comparable results.
- Possible outcomes - The once trained model can be used for multiple applications - Ad recommendation, Personalization, Fraud detection, Purchase propensity.

Concept Evaluation

Variable Sequence Length

It is intuitive that length of sessions can vary in terms of number of events made by the user. The proposed transformer based approach can handle variable session lengths.

- Nevertheless it is beneficial to eliminate very short sessions with meaningless interactions.
- A common solution in BERT based model is to add a padding token [PAD] by which they are left unattended and cause no influence.
- The only component that is length dependant is the position encoding which is not a part of transformer encoder therefore the variable session length should not impact the training performance.
- Dynamic batching approach by packing sessions of same length in one batch and padding the shorter ones to the maximum length of the session in that batch.

Bucket i	I	read	.	<PAD>			
	Je	lis	.	<EOS>			
			...				
...							
Bucket j	See	you	in	a	little	while	.
	A	tout	a	l'heure	<EOS>	<PAD>	<PAD>
			...				

Concept Evaluation

Influence of Vocabulary size

From the provided sample we can infer that the dictionary of distinct events is formed by combining the event type and value.

- The dictionary can be limited if we forecast the applications that would use the trained model for session encoding. If already known then apply:
  - Clustering of event type-value pair
  - Tf-idf based event type-value pair selection (Documents = User session, Terms = Session events)
  - Event importance ranking using Zipf method - Pattern2Vec

- BERT-Sentence transformer uses WordPiece tokenization. The vocabulary is constructed on the fly and is not equal to the number of distinct events in the dataset.

- Need to accept the fact - Due to sub-word based tokenization the vocabulary size might be exorbitant compared to event-level or char-level tokenizers. However, it is superior in handling OOV words.

- The performance of the model should not be impacted by the size of the vocabulary for a downstream task because the parameter in the vocabulary is updated only when it appears in the input text.

Concept Evaluation

Quality of User behaviour representation

The output of the training procedure produces a model which can be saved and imported for inference purpose to solve any downstream task.

Before inference, the trained model should be evaluated in terms of the quality of embeddings produced on a near to real time annotated test data.

Possible evaluation strategies:

- Perform dimensionality reduction using UMAP and clustering of user behaviour patterns and evaluate them using conditional entropy metrics.
- Visual interpretation of clusters using T-SNE plots on 2D or 3D space
- Rand Index, Precision and F1-Score metrics of clusters (In case of smaller dataset or larger # clusters)
  - $RI = \frac{TP + TN}{TP + FP + TN + FN}$, $P = \frac{TP}{TP + FP}$, $F1 = \frac{2TP}{2TP + FP + FN}$
- Annotate a hold-out data to indicate user intention and retrieve top-k similar users for a set of unseen test samples

Applications of Session Embeddings

1. Modelling Session Activity with Neural Embedding:

- Learn a latent manifold that captures users' session activity and can be utilized for contextual recommendations of apps in an online app store.
- User behavior modeling to predict purchases. One of the latest competition "Tmall Recommendation Prize" requires to predict future user purchases on Tmall website.
- Predict the next purchased item given a click action that made only in a predefined window before the purchase.
- Personalized marketting strategies by understanding the stage of the funnel they were in during the last interactions.

2. Personalization factors:

- Customer factors - Income level, Shopping process during information collection and purchase decision-making stages.
- Comodity values - Price and Quality
- Business Image - Credit evaluation, Quality of Service
- Willingness to shop - Content based filtering for personalized recommended products

THANKS!

Any questions?