Monday papers:

“Recurrent recommender networks (Wu et al., 2017)”

1. Since the user and movie representations are learned separately, I feel that the combined effect of user-item interactions are not captured. If there were links between the 2 LSTMs (something similar to attention models) representing the weight of how much one user’s feature is important for one movie’s feature, I feel that it would be a much better model.

2. Wouldn’t a bidirectional LSTM be more helpful because the model would have lot of inputs to learn the representations. Now, its limited because the future information cannot be reached from the current state. Also intuitively, the model would encode the temporal features more accurately than with a single-direction LSTM.

3. Is there someway to have a single LSTM (as training both the LSTMs could be computationally infeasible) that could take as input the concatenation of both the embedding matrices. Is there any drawback to that?

By Vignesh Gokul

1. **The author mentions that a movie’s impression tends to degrade initially in the first year after its release, and then increase with age, but the timespan of the test data is less than 1 year. I think it would be better to have the test data longer than 1 year.**

2. **Would it be helpful if the model also take the review texts as the input (e.g. word vector)?**

3. **Different users like different kind of movie, so I feel that adding the category of movies into embedding space would be helpful.**

(By Zeng Fan)

1. The authors state that RNN is more robust across different datasets as compared to PMF and SVD as the model learns the function that finds parameters instead of learning the parameters directly. How does this make the model more robust? Don’t the functions learnt also depend on the sparsity of datasets? What are the advantages/disadvantages of learning parameters (like in previous papers) directly vs the functions that learn parameters?

2. Does learning the user state RNN and movie state RNN separately leave out any correlations between user ratings and movie ratings? Are there any disadvantages of the subspace descent method?

3. What are some reasons why RNNs perform better on cold start problems as compared to other methods?

By Akanksha Grover

1. Considering the case of next POI/item/movie recommendation, would a evaluation metric like Hit Ratio and/or NDCG be more preferable over RMSEs for sequential modelling through RNNs? (Kiran Kannar)
2. The authors state that the retraining is not necessary with new data. Does this not mean that significant updations to the latent space do not happen i.e. the user and item latent spaces are not common, and are independent of each other? If not, why does new data and updation of current state mean no updations in latent space?(Kiran Kannar)

1. The stationary modules of the LSTM are pretrained with the models that are used as baselines. To me, this means that the backbone of the function that is being learnt are the parametric functions used in PMF and U-AutoRec.
2. The U-AutoRec is initialized with 500 hidden states. However, 500 was found to be the optimal number of hidden states for the 1M Movie Lens dataset. The authors of the RRN have not optimized the number of hidden states for the IMDB or Netflix papers whereas they have done extensive cross validation for their own models. This might make for a weaker baseline.

By Siddharth Dinesh

1. What is the advantage of finding the function which find parameters than find parameters? Does it mean that finding the function can avoid overfitting?
2. How can this model catch the seasonal effect? It used previous state to generate the next state. In this way, how does the authors can get the seasonal effects?

By Kuang Hsuan Lee

Authors say that they split their data based on time to simulate the actual situations where we need to predict future ratings instead of interpolating previous ratings. Any thoughts on following

1. Authors presented plots of rating vs time for awards, UI changes etc but did not provide the ground-truth ratings vs time. Are these plots of predicted rating (Fig 5-8) on the test data or the training data? This model can capture “age effect” but can it capture future events like awards, User Interface changes without re-training?

2. It seems like RNN in this model is a time-series autoencoder which tries to follow/track ratings of a movie. Seen a sudden increase in ratings (possibly due to an Award, UI change), it would increase the future predicted ratings. Certainly there would be a lag of some timesteps.
- Kulshreshth Dhiman

Neural Collaborative Filtering - He et al. WWW 2017

1. What is the hadamard product learning here? It says that the inspiration for that layer is taken from Neural Tensor Network, but NTN are primarily used to capture different kinds of semantic relationships. How is it becoming useful here (besides empirically improving the evaluation metrics)?
-Rahul Dubey

1. In collaborative filtering we had 1 latent space in which we place the users and items through the latent vectors. Do we have a similar analogy here? The FC embedding layers update their weights through backpropagation. Does the first neural FC layer backpropagating the gradients to the 2 user/item embed layers cause them to be in the same latent space? (Kiran Kannar)
2. Unlike collaborative filtering, the latent vectors are learnt through the embedding layer and X neural CF layers help in discovering the latent dimensions. Can this be elaborated intuitively on how each layer backprops this information so that the latent vectors are correctly identified? The user item interaction in this paper is the implicit feedback (1/0) The input are the identity one-hot vectors for each user and item. The output is 1/0 depending on whether the interaction was observed. (Kiran Kannar)
3. The authors compare GMF (Log loss/pointwise loss) with BPR and state that the pointwise loss is better than pairwise loss. Shouldn't the comparison be GMF(log loss) vs GMF (pairwise loss/BPR) ? BPR has been shown to be more effective than simple MF that uses pointwise loss. (See Rendle et.al) (Kiran Kannar)
4. The authors give the example of 4 users (u1 to u4) to show how inner product in collaborative filtering can place non-similar users closer in a ranking loss. How is this issue resolved using DNNs? If it's because of non-linearity, how does this help place them effectively in the latent space. (Kiran Kannar)

1. The authors mention that they only predict the interactions for 100 movies for each user from the dataset. If such a computationally expensive system has to be deployed in practice, what sacrifices could be made to allow predictions for the entire dataset for many users at a time?
2. Another way of looking at the above question is to first filter the most probable 100 movies (based on relevancy(?)) and use this neural network to predict the binary interaction scores for the selected 100 movies. How can this pre-selection be done to maximize the chances of 5-10 being good recommendations from the preselected 100 movies?

Siddharth Dinesh

1. The authors modeled implicit feedback from ratings which is clearly explicit feedback. Is this right approach ? a less rating clearly conveys negative user preference but lack of implicit feedback for a particular item doesn't mean negative correlation. So is it correct to just convert the ratings to binary ?(Sai Chaitanya Kolasani)
2. How does randomly assigning some of the sparse entries as negative improve model performance ? Doesn't this mean we are modelling something like the user may dislike some of the random items he didn't give implicit feedback ? (Sai Chaitanya Kolasani)
3. I think this approach suffer with the same cold start problem that occurs in traditional matrix factorization methods. Maybe we can augment the network with other auxiliary features to help alleviate this problem ? (Sai Chaitanya Kolasani)

1. Extension to include temporal behavior is not trivial,do we use the temporal information from the initial layers or is it just introduced in the end ?. Could this model be used to learn a time varying representation of users and items - using snapshots at various intervals? (Balasubramaniam Srinivasan)
2. Performing negative sampling and pointwise loss appears to be counter intuitive. Firstly negative sampling helps majorly when we definitely know an interaction to be of negative nature (and not of unknown nature - negative sampling trains based on sampling places it might have expected an interaction, but didn't find one). Therefore while these rate the positive examples better than negative examples - ordering within negative samples could be affected. (Balasubramaniam Srinivasan)
3. The author doesn't give details / compare the running times of the two. How big a factor is this? (Balasubramaniam Srinivasan)
4. There have papers about "Neural Word Embedding as Implicit Matrix Factorization" - based on the word2vec model. So if modelled well, would the neural techniques still surpass the matrix factorisation results? (Balasubramaniam Srinivasan)

1. **Dot Product :** The paper mentions an example with 2 dim vectors P1,P2,P3 and placing P4 according to similarity with P1 creates a misinterpretation when similarity is observed w.r.t P2 and P3. This is sighted as a shortcoming of dot product similarity. Is it correct? By just assuming a 3D vector representation for all these vectors we can imagine that the dot product similarity will overcome this limitation where P4 can still be equally inclined to P3 and P4 (as it should) when placed appropriately at an angle with P1. thus my question is regarding the crux of the paper, can you sight any example where the dot product similarity can lead to incorrect notion of similarities. - **Dhruv Sharma**
2. The paper tries to model the interaction between the latent representations of user and item using a neural network. This is a replacement of the dot product between gamma (u) and gamma (i) in the MF objective function. But we have seen Beta(u), Beta(i) and temporal terms as well in the various MF objective functions. How do you think the network would be extended to incorporate these? - **Dhruv Sharma**
3. Do you think other MF methods such as PMF, NMF can also be general cases of NeuMF?

- Dhruv Sharma

1. **Why do the authors train network with sgd instead of mini-batch? If we use mini-batch, we can benefit from the paralyzing computer.**
2. **The vector is 0/1, I think using dot product is not best approach to measure the similarity, how about jaccard similarity?**
3. **Usually there are a lot of noise in Implicit data, can this model handle with this kind of noise and also, it seems it will suffer from cold-start problem, how to add the context-based in this model?**
   Kuang Hsuan Lee

Sequential User-based Recurrent Neural Network Recommendations (Donkers et al., 2017)

1. The Attentional User base GRU works better for the MovieLens dataset as compared to the Rectified Linear User Integration but does this performance depend on the dataset? Is there an advantage to control user input to GRU vs controlling both item and user inputs?

2. Will the Gated Recurrent Unit have better performance than LSTM since here both user and item vectors are together controlled as inputs to model whereas there were learned separately in LSTM.

By Akanksha Grover

1. Instead of modifying the GRU unit to take both user and item vectors, why not concatenate and pass them to GRU? What's the advantage of one approach over the other?

- Rahul Dubey

Collaborative Variational Autoencoder for Recommender Systems (Xiaopeng Lie and James She., KDD, 2017)

1. The authors suggest a denoising interpretation for the proposed model with the inference network "corrupting" latent content variable with some noise. I am not sure I understand this interpretation, since from my understanding, the inference network is used to learn a distribution over the item content variables in their latent space.
2. Temporal trends

Chester Holtz

TransNets: Learning to Transform for Recommendation (Catherine and Cohen, 2017)

3. The authors suggest to use the output of the dropout layer for the source network to match the output of the CNNTextProcessor from the target network as it is regularized. Why is the choice not reflected in the target network? Could they have used the output of the dropout layer from the target network and source network? (Dhruv Sharma)
4. The paper requires the <user,item> with both ratings and reviews. It is not common for people to add both ratings and reviews for a product. They might add reviews but not ratings (hardly happens I think) or add ratings but not reviews (is more common). They <user,item> will never be used for learning the representations. Could they have used implicit feedback which is more readily available? (Dhruv Sharma)
5. Also, some people tend to write reviews only if they did not like the product (low rating) and some write for only the products they liked (high ratings). Would the dataset be biased towards ratings on both ends? (Dhruv Sharma)
6. They did not compare the model with similarity based networks such as Siamese model. Do you suppose it could have served as a baseline to evaluate against for this problem? (Dhruv Sharma)

1. Usually reviews are written to highlight unique aspects of the places you visit, unless the review form has a structure to it such as in the Beer Reviews dataset. I think further preprocessing of reviews to identify and separate out aspects of reviews might be an interesting extension to the model. Would this already be captured by the neural network model? (Siddharth Dinesh)

1. As mentioned above, since its not necessary that every user should write a review, we might end up not having reviews during the training process. Would a semi-supervised approach be useful in this case? (By Vignesh Gokul)

How does the approach discussed in the paper, with multiple networks joined by a Factorization Machine, compare to models that only use one of these approaches? E.g. how is the performance if we rely completely on the neural net to make recommendations? (Rajiv Pasricha)

How well does this method scale to larger datasets? The Yelp and Amazon datasets are pretty standard in size but is the network able to effectively train on a much larger datasets? (Rajiv Pasricha)

Using a neural network-based approach, is it easier to add additional relevant sources of data to the model? How would we add data such as implicit feedback or social connections? (Rajiv Pasricha)

1. On yelp dataset, there are 20% reviews are fake reviews, does this model tolerate the noise or the fake reviews?
2. If the dataset is very sparse, does this model work?
3. Could this model apply for on-line, which means when there is the new data coming, will this model update the trained network or it directly uses the pre-trained network? Kuang Hsuan Lee

---

Wednesday Papers:

"Deep Neural Networks for YouTube Recommendations (Covington et al., 2016)"

I think if someone watch the same video again and again, it will show strong signal, how to catch this signal? I mean in the short time gap, the user watches some video many times. Does it use threshold to transfer to binary variable or just count? If just count, how does it decide the time gap?
Also, does it transfer the model this situation? If we know the user watches A after watching B, and the time gap the user watches them is 2 hours,
In scenario1, A, B is music, it means after the user watch video A, the user exit or do other things, and after near 2 hours the user watches video B.
In scenario2, A, B is TV shows or longer videos, it means after the user watches video A, the user directly watches video B.
Does it consider the difference in this case? How to give this case as the input to Neural network?
Youtube has many playlist from users, but it seems the paper does not consider that feature, if we would like to consider the feature, how to change the model?

By Kuang Hsuan Lee

1. The paper suggests watch time as a better way to represent proclivity towards watching a video completely against explicit feedbacks such as clicks,likes etc which may be more sparse. Do you think it may not convey correctly for videos of different lengths? A small video is more likely to be watched (and still disliked) than say a 2 hour length video which the user watched for say 1.5 hours? *Dhruv Sharma*
2. The paper says that they use a function of expected watch using A/B testing for evaluating loss. Since this is eventually a personalized ranking model, do you think a BPR optimization would have worked on their scale? *Dhruv Sharma*
3. In sites like youtube and twitter there are often surges in views of videos when released, although they do not match everyone's "taste". Do you think the candidate generation algorithm, as explained in the paper, explains these candidates generated powered by global peak in views? (Dhruv Sharma)

[Stephanie Chen] Could you explain more about the "example age" feature in section 3.3? The authors describe correcting for a model's implicit bias towards the past, because they're trained on datapoints that occurred in the past. How is this fixed? Also, what is cross-entropy loss (referred to in section 4.2)?

1. The paper doesn't mention about explicit feedback from the user such as channel subscription, likes which may greatly affect the candidate generation. Even though explicit feedback is sparse, but using it may improve the recommendation. Is it still advisable not to use it?
2. It may very well happen the recommendation may be of periodic nature. Based on time of day or day of week, the recommendation should vary. Would including time factor explicitly during candidate generation phase help in recommendation?
3. The paper mentions that we take equal datapoint from each user set to prevent domination of loss. Shouldn't we cluster the user of similar interests and weight all the user cluster equally? By a cluster, I mean users tending to have similar interest.
4. We are using language models for embedding videos and then these embeddings are passed as parameter to the network. Is it possible to give raw input to the network and let the network learn embeddings itself?

By Sudhanshu Bahety

1. While taking into account search token based embedding, shouldn't we also factor in the time when the search was made. As recent search should be given more weightage than past searches?
2. Recommending videos based on a sequence will always tend to rank videos of very similar background together thereby reducing diversity in result. How is it possible to incorporate diversity in the recommendation so that user may explore different kinds of videos as well?
3. Paper mentions that predicting future watches are better than prediction held-out watches. But assuming we use held-out watches, how can we use to predict the candidate generation during serving phase as we won't be having the future inputs to the network?

By Digvijay Karamchandani

1. Since the watch history is a time series data, wouldnt a recurrent network be more helpful in predicting the next watch. The input at each timestep would be the concatenated embeddings of the video,search etc as described in the paper.

2. I believe the author samples positive examples by checking if an user completed watching a video. This is a vague description as people tend to ignore the last few seconds of a video and just move on to another one often. Either we can have a particular time limit for each user. Once an user watches a particular video for that time limit, it can be treated as a positive example. Or we can move to some features like the likes made by an user or something like that.
3. The model learns the embeddings of the watches based on a Continuos Bag of Words model. Since skipgram model learns more finer features if there is a lot of data, I believe this model could be improved if the embeddings were learned in a skipgram approach.

By Vignesh Gokul

1. What is the motivation for using extreme multiclass classification? Why would not a logistic loss function or a pairwise ranking work?
2. Authors average the embeddings of user's watch history (sequence of videoIDs). What does this average represent in the embedded space?
   A user can have diverse interests (say different genre movies/songs etc) and watch history can contains videos of totally different content. Is it a good idea to average all this information to a single point in the embedded space?

3. How does setting “age” feature to zero during serving time boosts recent videos?
4. If videos are watched through other means, can this be personalized as users are no longer Youtube users?
5. Authors say if a users did not watch recently recommended video then this impression is demoted in next page load. Since we are making top-N recommendation, user would choose only one of them but this doesn’t mean demoting other N-1 videos as they are still relevant. How does this work?
6. Is this model good for playlist prediction? Playlist prediction would require recommending different content (not too similar content) sequentially.

- Kulshreshth Dhiman

1. Why and how do the authors sample negative classes from background distribution?
2. The information of a new updated videos may change rapidly (e.g. number of clicks, number of comments), so the embedding of these videos can be not stable. How do the model learn the initial embedding of the new updated videos?
3. What’s the relationship between the input video embedding and the output video vectors?

(Zeng Fan)

1. The paper mentions that it is challenging to represent a temporal sequence of user actions during feature engineering. Will it make sense to come up with general “whitelist" user activity patterns on videos and then matching them to user activity sequences and further, getting features out of it?

2. While modelling expected watch time, can further adjustment of the weights by considering additional factors like thumbs up/down, comment, channel subscription after watching video, etc. improve the performance?

3. Predicting "next watch" resulted in better performance over holding out a randomly held-out watch. Can a hybrid model using both techniques result in even better performance, by accounting for different kinds of favourable watch patterns?

By Nitin Kalra

1.

1. Sometimes a user may click a video by mistake and as such the duration for which a video is played is very short. Does it take any correlation between the click and the duration into account in order to identify mistakenly clicked videos? Or it does include that mistaken click in its recommendation.?
2. One big advantage of using neural nets is that you don’t have to do the manual feature engineering. But here they do that. Is it possible to use a neural network separately to get the user/item features automatically?

3. Does it take into account the videos user has liked or added to the playlist?

By Rishab Gulati

Neural Factorization Machines for Sparse Predictive Analytics (He and Chua, 2017)

1. How do you think NFM can combine with field-aware factorization machines, associate several embedding vectors for a feature to differ interactions with other features in another field?
2. Paper mentions that authors sampled 2 negative instances to pair with one positive instance, why it was able to ensure the generalization and how exactly it was implemented?
3. For sequential data, would the replacement of FNN with RNN may yield better performance?

By Yiwen Gong

[Stephanie Chen] What are your thoughts on the dropout regularization technique described in section 3.2.1 to prevent overfitting? They talk about it being effective in section 4.2.1, but it seemed pretty complex: randomly drop some neurons and their connections, but only apply dropout at certain times and to certain layers of the neural network, etc. So, an explanation about how it works, and its efficacy would be great.

1. Eq 2: Why do you need to separately model linear regression part? Can't f(x) learn this automatically?
2. They convert each log (Frappe dataset) to a feature vector using one-hot encoding. How do you handle continuous context variables?
3. They did not fine tune size and dropout ratio for each hidden layer separately. NFM-2 (2 hidden layers) doesn't improve performance over NFM-1. Can this be due to not selecting optimal hidden layers? Fine-tuning deep layers should improve performance.
4. What could be the reason for pre-training not working for NFM? This means that embedding learnt during training and the one pre-trained using FM is significantly different.

- Kulshreshth Dhiman

What Your Images Reveal: Exploiting Visual Contents for Point-of-Interest Recommendation (Wang et al., 2017)

1. Sequential/Temporal information has been sighted on multiple occasions as positive contributors to POI recommendation problem. How do you think the paper can be extended to include these? If more matrices are added for the same do you think the model will grow computationally (**Dhruv Sharma)**
2. The paper uses 2 major interactions <User, Images> and <Location, Images> to model the recommendation. The model also suggests to remove “Selfies” from the data as they reveal little about the features of the sites. Do you think the location associated with such images can be used as a third matrix <User, Location>? Do you think this can help for a class of cold start users? (**Dhruv Sharma)**
3. The model is only evaluated against other MF methods. Do you think that serves a correct comparison? Since there is image data involved, would an end-end NN technique like Siamese similarity be considered ? In your opinion will the method outperform in such a comparison (**Dhruv Sharma)**

1. **Do you think if we add the description into deep learning to get the features and also put them into the model, does it benefit also?**
2. **Do you think if we classify the image to the locations it tag in neural network instead of getting the features first, is this method better? If not, what is the disadvantage for this method?**
3. **The author use whether or not the user visit the location, sometimes the user would not like the location, do you think maybe we could use the number of pictures taking in the location to measure the love of the location. For example, user A prefer location U than location V, the user A takes 100 pictures in location U and take only 10 pictures in location V, in this way, the target value is real value instead of boolean, also we can normalize the value to 0 - 1 floating number. Do you think this setting will increase the performance?**
   Kuang Hsuan Lee

1. **The paper mentions that less than 30% images are explicitly tagged with locations. Do you not think this will make the training, and subsequently the predictions, skewed?**
2. **The author mentions that VBPR and VPOI show significantly less performance degradation compared to other methods. This has been justified for VPOI. Why VBPR?**
   Aditi Ashutosh Mavalankar

1. The weights of the CNN are updated during the training process to learn image features relevant to the task of predicting user POIs. Is it possible to interpret these new weights in order to figure out which image features are useful predictors?
2. It appears that there are many assumptions being made in the structure of the model - in particular the Gaussian priors for the parameters or the negative sampling optimization procedure. Are these assumptions valid? How would we be able to test that?

3. How would VGG16 compare to other CNN architectures at this task? E.g. ResNet or AlexNet, which have also been trained on ImageNet and show good performance.

Rajiv Pasricha

3D Convolutional Networks for Session-based Recommendation with Content Features (Tuan and Phuong, 2017)

1. The author uses max pooling for three type of features: item_id, item description and category. Does this model include the interaction between them?
2. Do you think we could add image information to this model to improve the performance?
3. Why do we use character-level encoding? What is the advantage of this method compared with bag-of-word encoding?

   Kuang Hsuan Lee

1. Architecture - How is it beneficial to use 3D CNN to model sequential data (user clicks) over something more conventional (like RNN, as in the related work) for sequential data? Do you think that this is a conscious design choice that works better for this use case, or is it something that just performed better (trial and error)?
2. Feature Representation - The authors claim that their model has a reduced number of parameters due to the use of character-encoding over 1 hot encoding of features. How can this method be adapted for features that are not composed of text or characters? Also, they decrease size by ignoring descriptive text/data. Is this feasible in general or just suited to this dataset?
3. Results - Considering the small margin by which 3D CNN beats state-of-the-art models (only when additional features are considered and when lesser number of items are considered for recall), would it be fair to say that the model is perhaps tailored to do well on the chosen datasets and may not beat state-of-the-art models on other datasets?

- Shreyas Udupa

1. While preparing training data set, will it be a good idea to remove cases when a product was added to cart, but subsequently removed? Such cases are not very uncommon. Will such a change add any value?

2. The paper suggests that in a long session, first and last clicks are more important over the middle ones. Instead of blindly removing any particular segment of click stream

as a whole, how about weighing the importance of each click (e.g. time spent viewing or learning more about each product) and construct a trimmed stream of clicks with the ones that would be most influential, although still maintaining the chronological ordering?

3. While creating character-level representations, the paper suggests that additional description is ignored to reduce model size. Instead, will it be a good idea to process the description and come up with a number of most significant words in description and concatenate them to the name of product? Will it lead to any significant improvement in practice?

By Nitin Kalra

1. The input sequences are preprocessed by removing some of the middle events and the authors explain they seem to be irrelevant. This seems a reasonable assumption for small/midsize vendors with a small inventory but for websites like amazon where this type of recommendation task is very helpful this approach may fail. Most of the times due to the sheer size of available models users browse products randomly searching for good deals. So a scheme which weights the relevance of each click to the session seems more appropriate.(Sai Chaitanya Kolasani)

2. The proposed 3-d CNN architecture seems to handle only fixed length sequences is there a way to handle variable length sequences ? (Sai Chaitanya Kolasani)

3. There are a number of magic numbers in the proposed model like the size of features chosen, number of convolutions etc., Can these parameters be tuned by cross validation ? Generally deep learning papers doesn't explain the architecture choices as tuning these many parameters by cross validation is computationally infeasible but these numbers seem arbitrary. (Sai Chaitanya Kolasani)

4. CNN's achieve good results for images because they have translation invariance. That is a particular spatial feature can be anywhere in the image and it will be identified by the conv layer because of the weight sharing. Assuming same mechanism in the time dimension how does 3-d CNN's capture temporal differences as the author claims ? If at all they should kind of ignore temporal differences because of the weight sharing.(Sai Chaitanya Kolasani)

Deep Learning based Large Scale Visual Recognition and Search for E-Commerce (Shankar et al., 2017)

1. I was wondering if we could also add user personalization to the visual similarity objective? Maybe giving weightage to that would lead to also recommending products that may be a little different in design(less visually similar) but also give preference to what the user likes.

2. I was curious about the fact that they got better results with wild image triplets as compared to using the object detection and localization network that provides bounding boxes in wild images.

By Akanksha Grover

1. During the training step, when they use the cropped wild image as the query image, why don't they use wild image as positive instead of ground truth from catalog, as not all sellers will post a catalog-like image for their products?
2. What would be the advantage to use euclidean distance for images feature vectors over other methods like cosine similarity or pearson correlation?
3. The results may also be improved by the clicks after people saw all these similar results.
4. I am confused that as the throughput of VGG+Shallow is only 118 QPS, how could the system support 2000 QPS? (Yiwen Gong)

1. Say there is a shirt with the Joker's face (Batman). We would ideally want similar shirts with Joker on them or perhaps Batman related aspects.
    a. The first aspect of Joker- face recognition within a shirt is a low-level **meta-object recognition** task here. Does the neural network actually capture this info?
    b. If so, can we transfer this to finding similar, yet complementary items across **category**?- say a shirt with Joker on it, and the trademark hat of Joker? Use the low level, finer details learnt in the shallow convolution layers and find similar abstractions across category? (Kiran Kannar)
2. Does the neural network distinguish the pattern as a print in the shirt and not the shirt stripe type? The TShirt has a lot of stripes. (Kiran Kannar)

 (query image)  **prediction**?

Vs. **desired prediction**  or 

3. The authors state that LSH and Kd trees did not product required in-production performance gain for k-NN search across high-dimensional embeddings. So they resort to full nearest neighbor search across the entire catalog space. What design decisions do you think led them to just brute force over entire catalog space, but handle kNN efficiently with incremental, streaming MR updates? (Kiran Kannar)

1. Is there any particular reason for using the VGG-16 network, instead of the ResNet-50, which is theoretically faster and better in performance?
2. The throughput of VGG-16 for this task is only 5 qps using CPU. Is this too small for realistic scenarios? Are there any improvement suggestions for the efficiency of this work?
3. For results on the Pants Category, why is VisNet-FRCNN better than VisNet (38.5 vs 31.8)? Is this due to the inaccurate cropping of RoIs in the training set? (Siyu Jiang)

1. If we can take implicit feedback from user, then wouldn't it capture the visual similarity as well because people would tend to click on similar items before purchase or for comparison? Also, apart from addressing cold start issue, is there anything else this network will have a significant impact on?
2. For generating training data we use BISS. How do we ensure that network won't learn the BISS function itself?
3. In related work, author mentions Siamese is unable to learn fine-grained features, but doesn't that depends on how you are training your siamese network?

-Sudhanshu Bahety

1. We use different network for different categories. Can't we increase the depth of the network and use a single network for all the categories?
2. In result, VISNET-FRCNN performs worse than VISNET although the former is fine-tuned on wild object. Shouldn't it perform better?
3. If we are using metadata for extraction, it should capture the similarity to some extent. Shouldn't the author have a baseline taking only metadata and implicit feedback into account and compare it with VISNET or is it guaranteed VISNET will perform better?

-Digvijay Karamchandani