1. Motivation: Recommending Videos to Users in News Service

Task:
- Given: Video & News services
- Goal: design a Recommender System (RS) that suggests videos to users who
  A. Have never used Video service before
  B. But used News service
- Constraint: Few/No users shared across services

Use case:
- News = popular & having a large user base
- Video = less known (e.g. relatively new service)

Making quality recommendations attract new users in News service who have never used Video service

Challenge: Conventional RSs don’t work

- Learning from Video users
  - Optimised for video users (input = video histories)
  - News users don’t have video histories
- Learning from News users ≠ possible (no labels)
- Learning from shared users ≠ feasible
  - Few/no shared users
  - Not enough training examples

Q. How should we utilise knowledge of Video service and transfer it to News users?

Our Approach:
- Adapt Video RS to News users with extreme classification + domain adaptation

Contribution:
- Propose a method that works with commonly available forms of content information

2. Problem Formulation

Recommendation as Extreme Classification:

Given:
- Video data (source) $D_S = \{x_S^1, x_S^2, \ldots, x_S^N\}$ i.i.d. $P_S(X, Y)$
- News data (target) $D_T = \{x_T^1, x_T^2, \ldots, x_T^M\}$ i.i.d. $P_T(Y)$

\[
\begin{align*}
   &x_S^i \in \mathbb{R}^d, \text{ vector representing, } x_T^j \in \mathbb{R}^d, \text{ vector representing video history} \\
   &y_c \in \{1, \ldots, K\}, \text{ video label}
\end{align*}
\]

Goal = Construct a classifier $\eta : x^i \mapsto y \in \{1, \ldots, K\}$ (predict a video corresponding to a news user)
with a low expected error: $Pr_{(X,Y)}[\eta(X) \neq Y]$

Note: Data domains are distinct $P_T(Y) \neq P_S(Y)$

- Supervised ML + Training on $D_S$ won’t work (error low)
- Correction via domain adaptation

3. Unsupervised Domain Adaptation

We use Domain Separation Network (DSN) [1]

DSN achieves domain adaptation with:
- Shared encoder: extract predictive features shared across domains
- Private encoder: extract features private to each data domain
- Shared decoder: reconstruct from private + shared features
- Objective: $L_{DSN} = L_{class} + L_{encoder} + L_{decoder}$

Fig: Architecture of DSN [1]

4. Experiment

Data = Video/News services of Yahoo! JAPAN
- 2 week-long browsing logs
- Training data = size 11m for each service
- No shared users in this data
- Validation data (33k) & Test data (38k):
  - Constructed from logs of shared users
  - Instance = (news history, video label) pair

Items have textual attributes:
- Video: title, cast, category, short description
- News: title, category

Note: only news articles in entertainment categories were used

Data representation:
- User history = bag of items
  - Treat as a document composed of item's textual attributes
  - Represent history with TF-IDF:
    - For each domain, form a vocabulary set according to TF-IDF value (computed from histories)
    - Combining two vocabulary sets
      → common vocabulary set of size 50k
    → Input dimension, $d = 50,000$

5. Discussions and Future Work

Discussion: Poor Performance of NN/DSN (CEL)
- Worse than POP, does not capture popularity
- Top-1 item prediction is too hard
- CEL does not give useful signal
- DCG better captures quality of predictions (given in the form of probability distribution)

Future Work:
- Replacing the training loss with a ranking loss (e.g. DCG)
- Combining item side information (info unique to RSs) using zero/few-shot learning techniques
  → ease the difficulty of extreme classification

References: