The multi-sources learning issue

- **Aim**: Increasing the diversity and generality of learned representations.
- **How**: By aggregating various data sources.
- **But**: Learned representations tend to be source-specific rather than multi-source.

**Source aggregated Dataset**

| S1: Star Wars | S2: Superman |
| S3: Indiana Jones | S4: Blade Runner |

**Targeted datasets**

- **4 Datasets**: Amazon Books (AB), Movies and TV (AM), Electronics (AE), Yelp challenge dataset (Yelp).
- **3 targeted datasets [2]**:
  - **Kept set (K-set)**: leakage between the source and the polarity.
  - **Rejected set (R-set)**: inverse leakage (w x t K-set) between the source and the polarity.
  - **Unseen set (U-set)**: sources not present in K-set and R-set.

**Procedure**


**Defining SCBow**

- Embeds word sequence $x = (w_1, ..., w_T)$ in $v_x = \mathbf{W}_v w_x$, where $v$ are word embeddings [1], $p(v|x)$ is a softmax layer,
- $v$ and the softmax weights ($\theta_{SCB}$) are learned with SGD minimizing $L_{SCB} = E_{x,y}[− \log p(y|x; \theta_{SCB})]$.

**Best results** are obtained with bigrams.

**Training details**

- $\alpha = 0.001$, $\beta = 0.05$
- Pretrain $v$, $\theta_{SCB}$ during 10 epochs.
- During 10 epochs do:
  - For a given mini-batch of data $B$, $v, \theta_{SCB}$:
    - $L_{SCB}$ = $−\alpha E_{x,y}[− \log p(y|x; \theta_{SCB})]$
    - $−\lambda E_{x,y,T} \log p(y|x; \theta_{SCB}) − \lambda E_{x,y,T} \log p(y|x; \theta_{SCB})$.
    - $\delta v, \delta \theta_{SCB} = −\beta E_{x,y,T} \mathcal{L}(v, \theta_{SCB})$.
  - Update $\alpha$ and $\beta$ with Adam.

**Regularizing SCBow with adversarial embeddings**

- $L = L_{SCB} + \lambda L_{id}$
- $L_{id}$ quantifies the source identifiability of hidden representations $v_x = (v_1, ..., v_T)$.
- Following the works [3, 4], we suggest an adversarial framework where we learn two embeddings of $x$, $v_x$ for the classification task and $\hat{v}_x$ for the source identification task.

**Training**

- Mean accuracy on K-set/K-set: $\lambda = 1$.

**Acknowledgements**

We thank Sidetrade for funding and for providing challenging and insightful datasets. We thank Pablo Pérez Antin and Jean-Cyril Schütz for fruitful discussions.

**References**