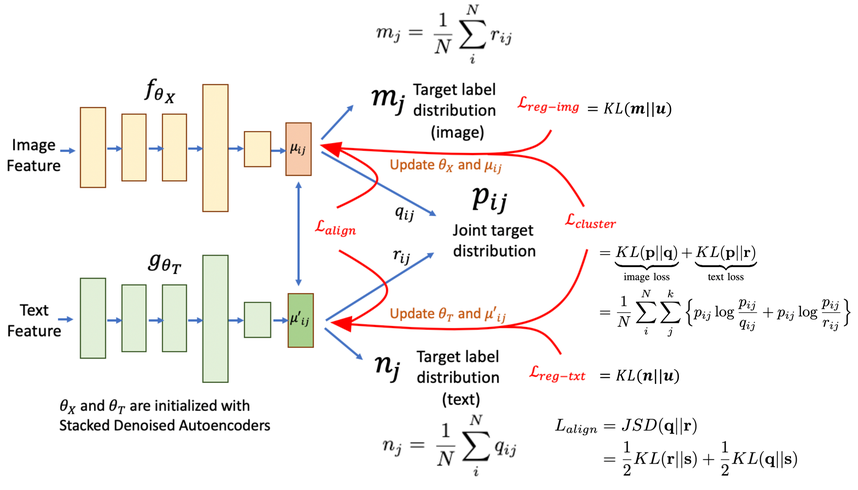
JECL Joint Embedding and Cluster Learning for Image-Text Pairs

Sean T. Yang (ttyang38@uw.edu), Kuan-Hao Huang (khhuang@cs.ucla.edu), Bill Howe (billhowe@cs.Washington.edu)

We propose JECL, a method for clustering image-caption pairs by training parallel encoders with regularized clustering and alignment objectives, simultaneously learning both representations and cluster assignments. JECL trains by minimizing the Kullback-Leibler divergence between the distribution of the images and text to that of a combined joint target distribution and optimizing the Jensen-Shannon divergence between the soft cluster assignments of the images and text. Regularizers are also applied to JECL to prevent trivial solutions. Experiments show that JECL outperforms both single-view and multi-view methods on large benchmark image-caption datasets and is remarkably robust to missing captions and varying data sizes.

Model Overview



Parameter Initialization: We initialize DNN parameters θ_X and θ_T with two stacked denoising autoencoders. We apply K-means to the initial embeddings to obtain initialized centroid set, μ_j and μ'_j.

Soft Assignment: We model the probability of data point i being assigned to cluster j using the Student's t-distribution, producing a distribution $\boldsymbol{q_{ij}}$ for images and $\boldsymbol{r_{ij}}$ for text)

Cluster Alignment: We use the Hungarian algorithm to obtain the alignment between image clusters and text clusters.

Joint Target Distribution ($\mathbf{p_{ij}}$): The joint target distribution aims to improve cluster purity and to emphasize data points with high assignment confidence.

$$p_{ij} = \lambda \times \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{j'} q_{ij'}^2/\sum_i q_{ij'}} + (1-\lambda) \times \frac{r_{ij}^2/\sum_i r_{ij}}{\sum_{j'} r_{ij'}^2/\sum_i r_{ij'}}$$

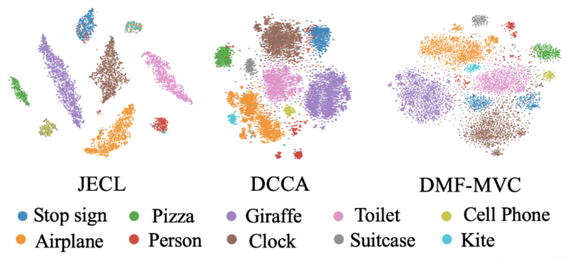
Overall Loss Function:

$$L_{JECL} = L_{cluster} + \gamma L_{align} + \beta L_{reg}$$

Results

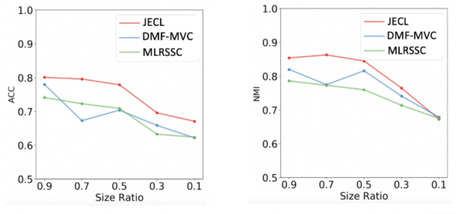
JECL outperforms the state-of-the-art multi-view clustering and image-text representation learning models on benchmark datasets by significant margins. The ablation study also shows both distribution regularizer and distribution alignment improve the overall performance.

	Coco-cross			Coco-all			Pascal			RGB-D		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Single-View (Image)												
ResNet-50 + KM	0.647	0.712	0.558	0.519	0.614	0.442	0.486	0.516	0.307	0.353	0.289	0.161
ResNet-50 + DEC	0.649	0.629	0.670	0.472	0.701	0.429	0.418	0.564	0.311	0.421	0.352	0.236
Single-View (Text)												
Doc2Vec + KM	0.720	0.852	0.737	0.613	0.807	0.589	**0.544**	0.602	0.398	0.438	0.384	0.279
Doc2Vec + DEC	0.720	0.843	0.729	0.557	0.738	0.501	0.295	0.294	0.120	0.429	0.383	0.287
Concatenation of Both Views + Single-View Models												
Concat(ResNet50+Doc2Vec) + KM	0.636	0.711	0.550	0.517	0.617	0.439	0.478	0.517	0.302	0.355	0.290	0.211
Concat(ResNet50+Doc2Vec) + DEC	0.737	0.758	0.677	0.419	0.550	0.275	0.225	0.326	0.121	0.344	0.255	0.172
Multi-View Representation Learning												
VSE + KM	0.665	0.736	0.607	0.520	0.628	0.430	0.479	0.508	0.300	0.388	0.318	0.194
DCCA + KM	0.712	0.822	0.703	0.645	**0.817**	0.603	0.442	0.485	0.238	0.388	0.310	0.186
CCAL-L_{rank} + KM	0.699	0.806	0.689	0.641	0.812	0.587	0.446	0.489	0.224	0.404	0.316	0.196
Multi-View Clustering												
BMVC	0.365	0.227	0.200	0.410	0.441	0.316	0.392	0.378	0.214	0.207	0.088	0.047
MultiViewLRSSC	0.726	0.781	0.706	0.569	0.747	0.530	0.534	0.574	0.371	0.474	0.400	0.277
DMF-MVC	0.829	0.805	0.774	0.632	0.776	0.608	0.512	0.573	0.380	0.441	0.330	0.257
JECL	**0.929**	**0.908**	**0.934**	**0.675**	0.801	**0.685**	0.512	**0.625**	**0.403**	**0.543**	**0.472**	**0.367**
w/o alignment	0.922	0.906	0.931	0.634	0.784	0.643	0.502	0.613	0.332	0.513	0.423	0.277
w/o regularizer	0.894	0.890	0.889	0.624	0.777	0.610	0.513	0.620	0.376	0.520	0.433	0.327
w/o alignment & regularizers	0.863	0.878	0.852	0.611	0.757	0.607	0.487	0.579	0.352	0.502	0.413	0.367



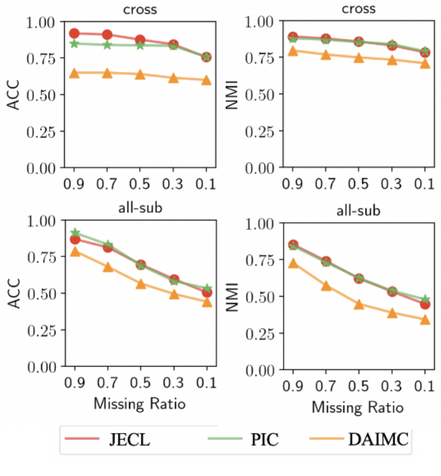
JECL successfully separates semantically distinct clusters with clear boundaries between clusters.

Robustness to Data size



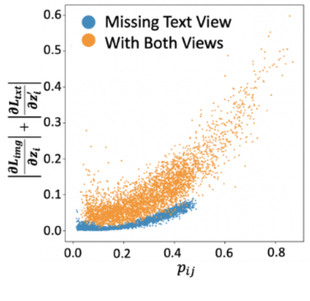
JECL performance as data size decreases. The performance degrades when size ratio is below 0.5(500 data points in each class), while JECL still outperforms the state-of-the-art multi-view clustering methods, DMF-MVC and MLRSSC on varying data sizes.

Robustness to Missing View



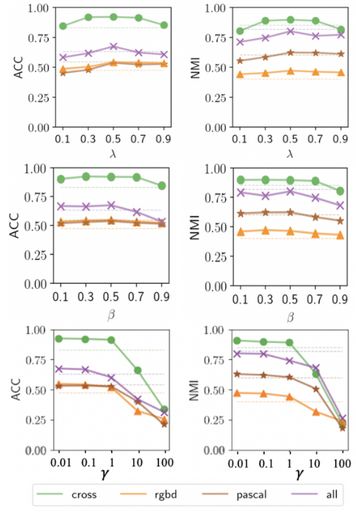
Experimental results on missing view scenarios. JECL is competitive with the state-of-the-art method, PIC, and outperforms DAIMC by a large margin on both datasets

JECL's robustness to missing data is attributable to the model of the joint distribution: the images with text (orange) contribute more to the gradient than the images with missing text (blue)



Robustness to Hyperparameters

JECL is generally robust to hyperparameter settings, while is the most stable and produces top results with $\lambda = 0.5, \beta = 0.1, and\ \gamma = 0.1$ among all datasets.



UNIVERSITY of WASHINGTON DataLab