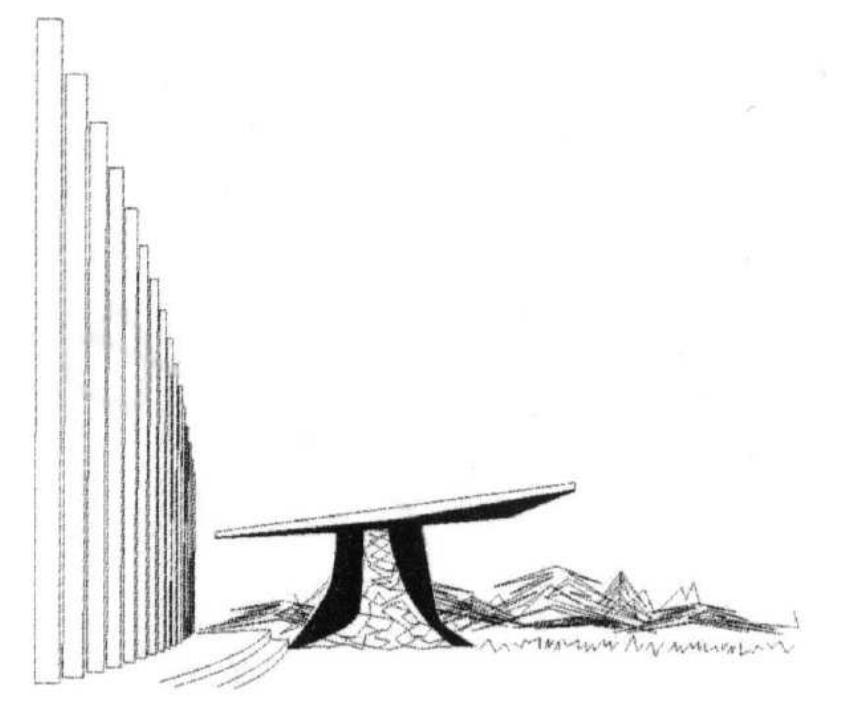


Wasserstein Barycenter Model Ensembling

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Cicero Dos Santos & Tom Sercu
IBM Research, Yorktown Heights, NY



Introduction

- Traditional ensembling methods: bagging, boosting, stacking, etc
- Popular ways to merge multiple models
 - Arithmetic averaging – rewards more confident models
 - Geometric averaging – rewards consensus across models
- What is missing?**
 - Ability to ensemble by incorporating side information of class relationships
 - Example: word prediction
 - p_i model's confidence on predicting word (represented as word embedding x_i)
 - $p = \sum_i p_i \delta_{x_i}$ - distribution over word embeddings/labels i
 - $C_{i,j}$ - dissimilarity between word i and word j
 - Ensemble by finding a balance between model confidence and label semantic similarity
 - Final ensemble can be strong even if models are confident on different (but semantically similar) words
 - Arithmetic/geometric mean cannot exploit this since they treat words as independent
- Wasserstein Barycenter**
 - Enables merging multiple probability distributions given a cost matrix $C_{i,j}$ between elements i and j
 - Balances model confidence and class semantic similarity

Model Ensembling

- Given m models, each defined by a prediction vector $\mu_\ell \in \mathbb{R}_+^{N_\ell}$, $\ell = 1, \dots, m$
- Goal: Find a consensus prediction $\bar{\mu} \in \mathbb{R}_+^M$
- Arithmetic mean $\bar{\mu}_a = \sum_{\ell=1}^m \lambda_\ell \mu_\ell$; Geometric mean $\bar{\mu}_g = \prod_{\ell=1}^m (\mu_\ell^{\lambda_\ell})$
- Wasserstein barycenter for model ensembling

$$\bar{\mu}_w = \arg \min_{\rho} \sum_{\ell=1}^m \lambda_\ell W_2^2(\rho, \mu_\ell)$$

Balanced W. Barycenter (normalized μ_ℓ)

$$\min_{\rho} \min_{\gamma_\ell \in \Pi(\mu_\ell, \rho)} \sum_{\ell=1}^m \lambda_\ell \langle C_\ell, \gamma_\ell \rangle$$

Unbalanced W. Barycenter (normalized μ_ℓ)

$$\min_{\rho} \min_{\gamma_\ell} \sum_{\ell=1}^m \lambda_\ell \left(\langle C_\ell, \gamma_\ell \rangle + \lambda \widetilde{KL}(\gamma_\ell \mathbf{1}_M, \mu_\ell) + \lambda \widetilde{KL}(\gamma_\ell^\top \mathbf{1}_{N_\ell}, \rho) \right)$$

Inputs: ε, C_ℓ ($|\text{source}| \times |\text{target}|$), λ_ℓ, μ_ℓ
Initialize $K_\ell = \exp(-C_\ell/\varepsilon), v_\ell \leftarrow \mathbf{1}_M$

for $\text{iter} = 1 \dots N$ **do**

$$u_\ell \leftarrow \frac{\mu_\ell}{K_\ell v_\ell}$$

$$p \leftarrow \exp \left(\sum_{\ell=1}^m \lambda_\ell \log(K_\ell^\top u_\ell) \right) u_\ell^{\lambda_\ell}$$

$$v_\ell \leftarrow \frac{p}{K_\ell^\top u_\ell}$$

end for

Output: p

Inputs: ε, C_ℓ ($|\text{source}| \times |\text{target}|$), $\lambda_\ell, \lambda, \mu_\ell$
Initialize $K_\ell = \exp(-C_\ell/\varepsilon), v_\ell \leftarrow \mathbf{1}_M$

for $\text{iter} = 1 \dots N$ **do**

$$u_\ell \leftarrow \left(\frac{\mu_\ell}{K_\ell v_\ell} \right)^{\frac{\lambda}{\lambda+\varepsilon}}$$

$$p \leftarrow \left(\sum_{\ell=1}^m \lambda_\ell (K_\ell^\top u_\ell)^{\frac{\varepsilon}{\lambda+\varepsilon}} \right)^{\frac{\lambda}{\varepsilon}}$$

$$v_\ell \leftarrow \left(\frac{p}{K_\ell^\top u_\ell} \right)^{\frac{\lambda}{\lambda+\varepsilon}}$$

end for

Output: p

Experiments

Attribute-based Classification

- Dataset: Animals and Attributes, 85 attributes, 50 classes
- 2 attribute-based classifiers
- Compared arithmetic/geometric means $p(c|\mu_{a,g}) = K\mu_{a,g}$ and Unbalanced Wasserstein Barycenter
- Similarity matrix $K \in \mathbb{R}^{50 \times 80}$

	Accuracy	resnet18 alone	resnet34 alone	Arithmetic	Geometric	W. Barycenter
Validation	0.7771	0.8280	0.8129	0.8123	0.8803	
Test	0.7714	0.8171	0.8071	0.8060	0.8680	

Multi-label Prediction

- Dataset: MSCOCO, 80 categories
- 8 classifiers
- Similarity matrix $K \in \mathbb{R}^{80 \times 80}$ based on GloVe/Word2Vec distances, word co-occurrences

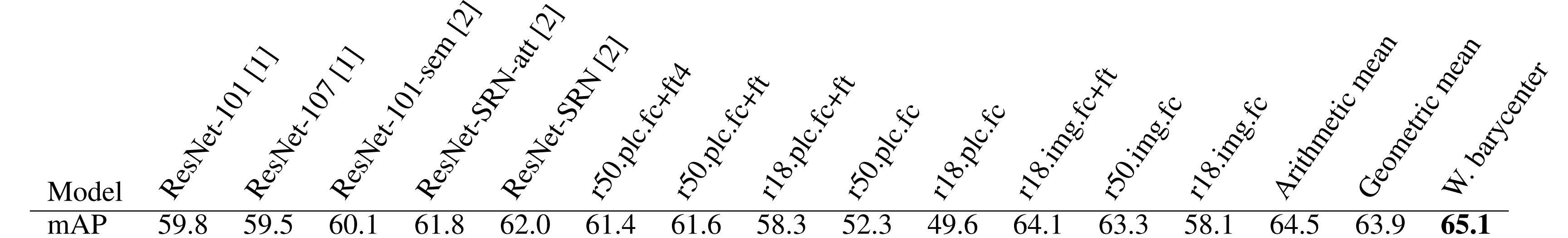
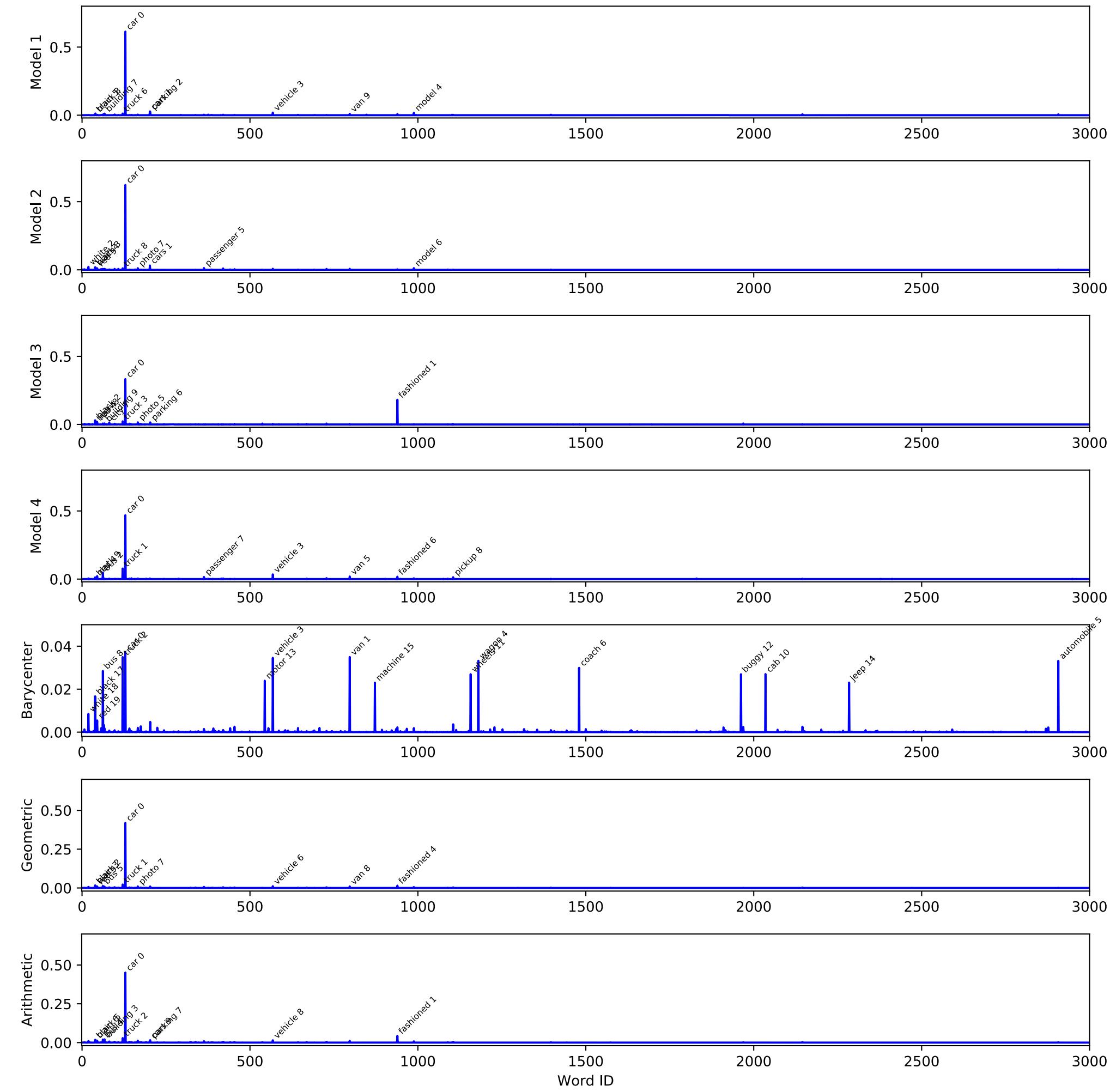


Image Captioning

- Dataset: MSCOCO
- 5 image captioners
- Similarity matrix $K \in \mathbb{R}^{10096 \times 10096}$ based on GloVe distances and word synonyms graph



Rank	W. Barycenter	Arithmetic	Geometric	Model 1	Model 2	Model 3	Model 4
0	car	03.73	car	45.11	car	61.37	car
1	van	03.50	fashion	04.37	truck	02.23	truck
2	truck	03.46	black	02.42	black	02.79	black
3	vehicle	03.46	buildin	02.10	train	01.51	vehicle
4	wagon	03.32	bus	02.00	fashion	01.49	model
5	automob	03.32	black	01.79	bus	01.30	train
6	coach	02.99	train	01.73	vehicle	01.14	truck
7	auto	02.98	parking	01.55	photo	01.01	model
8	bus	02.85	vehicle	01.49	van	01.01	black
9	sedan	02.71	cars	01.41	red	01.01	pickup
10	cab	02.70	photo	01.29	parking	00.94	train
11	wheels	02.70	red	01.26	buildin	00.88	black
12	buggy	02.70	van	01.26	vehicle	00.78	style
13	motor	02.39	white	01.04	passeng	00.81	model
14	jeep	02.31	passeng	00.92	white	00.67	time

