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 $\mu_\ell \in \mathbb{R}_+^{N_\ell},\ \ell = 1, \ldots, m$
- 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 $\bar{\mu}_g = \Pi_{\ell=1}^m (\mu_\ell^{\lambda_\ell})$
- 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)

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

Inputs: ε, C_ℓ ($|\text{source}| \times |\text{target}|$), λ_ℓ, μ_ℓ
Initialize $K_\ell = \exp(-C_\ell/\varepsilon), v_\ell \leftarrow 1_M$
for $iter = 1 \ldots N$ **do**
$u_\ell \leftarrow \frac{\mu_\ell}{K_\ell v_\ell}$
$p \leftarrow \exp\left(\sum_{\ell=1}^m \lambda_\ell \log\left(K_\ell^\top u_\ell\right)\right) u_\ell)^{\lambda_\ell}$
$v_\ell \leftarrow \frac{p}{K_\ell^\top u_\ell}$
end for
Output: p

Unbalanced W. Barycenter (normalized μ_ℓ)

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

Inputs: ε, C_ℓ ($|\text{source}| \times |\text{target}|$), λ_ℓ, λ, μ_ℓ
Initialize $K_\ell = \exp(-C_\ell/\varepsilon), v_\ell \leftarrow 1$
for $iter = 1 \ldots N$ **do**
$u_\ell \leftarrow \left(\frac{\mu_\ell}{K_\ell v_\ell}\right)^{\frac{\lambda}{\lambda+\varepsilon}}$
$p \leftarrow \left(\sum_{\ell=1}^m \lambda_\ell \left(K_\ell^\top u_\ell\right)^{\frac{\varepsilon}{\lambda+\varepsilon}}\right)^{\frac{\lambda+\varepsilon}{\varepsilon}}$
$v_\ell \leftarrow \left(\frac{p}{K_\ell^\top u_\ell}\right)^{\frac{\lambda}{\lambda+\varepsilon}}$
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

Model	ResNet-101 [1]	ResNet-107 [1]	ResNet-101-sem [2]	ResNet-SRN-att [2]	ResNet-SRN [2]	r50.plc.fc+ft4	r50.plc.fc+ft	r18.plc.fc+ft	r50.plc.fc	r18.plc.fc	r18.img.fc+ft	r50.img.fc	r18.img.fc	Arithmetic mean	Geometric mean	W. barycenter
mAP	59.8	59.5	60.1	61.8	62.0	61.4	61.6	58.3	52.3	49.6	64.1	63.3	58.1	64.5	63.9	**65.1**

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	41.94	car	61.37	car	62.25	car	33.25	car	46.88
1	van	03.50	fashion	04.37	truck	02.23	cars	02.79	cars	03.16	fashion	18.15	truck	07.74
2	truck	03.49	truck	02.92	black	01.67	parking	02.62	white	02.22	black	03.08	bus	04.78
3	vehicle	03.46	buildin	02.10	train	01.51	vehicle	01.93	black	01.95	truck	02.29	vehicle	03.46
4	wagon	03.32	bus	02.00	fashion	01.49	model	01.75	train	01.68	red	01.88	red	02.20
5	automob	03.32	black	01.79	bus	01.30	train	01.26	passeng	01.33	photo	01.57	van	01.93
6	coach	02.99	train	01.73	vehicle	01.14	truck	01.22	model	01.24	parking	01.52	fashion	01.74
7	auto	02.98	parking	01.55	photo	01.01	buildin	01.17	photo	01.21	city	01.41	passeng	01.56
8	bus	02.85	vehicle	01.49	van	01.01	black	01.04	truck	01.15	train	01.30	pickup	01.37
9	sedan	02.71	cars	01.41	red	01.01	van	01.04	red	01.15	buildin	00.74	black	01.29
10	cab	02.70	photo	01.29	parking	00.94	fashion	00.82	silver	01.03	fashion	00.72	train	00.79
11	wheels	02.70	red	01.26	buildin	00.88	suv	00.69	vehicle	00.78	bus	00.71	style	00.68
12	buggy	02.70	van	01.18	cars	00.81	automob	00.67	van	00.75	style	00.69	model	00.59
13	motor	02.39	white	01.04	passeng	00.71	parked	00.57	buildin	00.71	time	00.67	fire	00.57
14	jeep	02.31	passeng	00.92	white	00.67	picture	00.55	bus	00.70	old	00.58	white	00.52