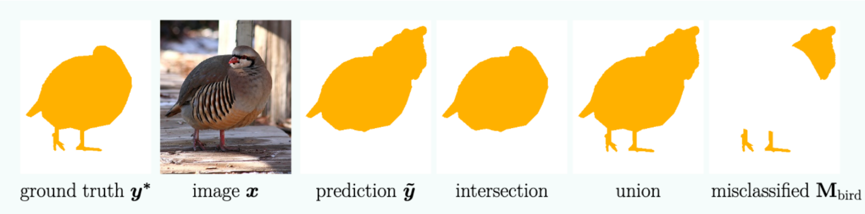
Optimizing Jaccard, Dice, and other measures for image segmentation

Matthew Blaschko
joint work with Jiaqian Yu, Maxim Berman, Amal Rannen Triki, Jeroen Bertels, Tom Eelbode, Dirk Vandermeulen, Frederik Maes, Raf Bisschops

KU LEUVEN
PSI processing speech & images

Motivation - Jaccard index



Jaccard = intersection/union = $\frac{|y \cap \tilde{y}|}{|y \cup \tilde{y}|}$

- No bias towards large objects, closer to human perception
- Popular accuracy measure (Pascal VOC, Cityscapes...)
- Multiclass setting: averaged accross classes (mIoU)
- **Function of the discrete values of all pixels**
  $\rightarrow$ Optimizing IoU is challenging!

Motivation - Dice score

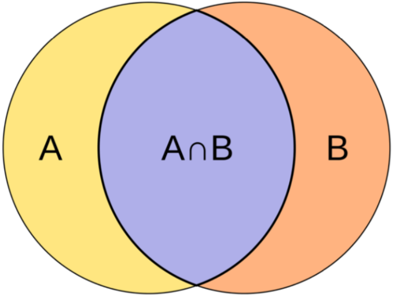
$\mathrm{Dice}(y, \tilde{y}) = \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|}$

- The *de facto* standard measure for medical image analysis
- Traced back to Zijdenbos et al., 1994
- Chosen due to class imbalance in white matter lesion segmentation
- Size and localization agreement
- More in line with perceptual quality compared to pixel-wise accuracy

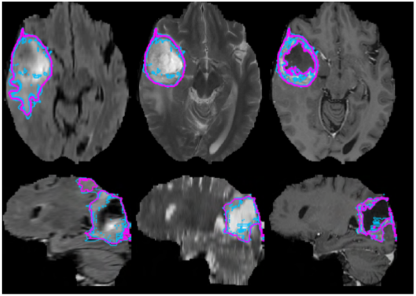
- A generation of radiologists trained reading articles reporting average Dice score

[Zijdenbos et al., IEEE-TMI 1994]

Jaccard & Dice



(a) Jaccard loss = $1 - \frac{|A \cap B|}{|A \cup B|}$



(b) Dice loss = $1 - \frac{2|A \cap B|}{|A| + |B|}$

Outline of the talk

- Similarities, LSHability, and supermodularity
- Jaccard & Dice measures
- Risk minimization
- Dice in the “real world”

Similarities

Definition (Similarity)

A function $S : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is called a similarity if

1. $S(X, X) = 1$;
2. $S(X, Y) = S(Y, X)$.

For a similarity S, the corresponding distance is simply $1 - S$.

LSHability

Definition (LSHability)

An LSH for a similarity function $S : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is a probability distribution $P_{\mathcal{H}}$ over a set $\mathcal{H}$ of hash functions definied on $\mathcal{X}$ such that $\mathbb{E}_{h \sim P_{\mathcal{H}}}[h(A) = h(B)] = S(A, B)$. A similarity S is LSHable if there is an LSH for S.

Proposition (Charikar, 2002)

If a similarity is LSHable, its corresponding distance is metric.

note: metric $\not\Longrightarrow$ LSHable

Supermodular similarity

Definition

A similarity S is said to be supermodular if, holding one argument fixed, the resulting set function of its symmetric difference $f_X : A \mapsto S(X, X \triangle A)$ satisfies the following conditions:

1. f_X supermodular;
2. monotonically decreasing, i.e. $f_X(A) \geq f_X(B)$ for all $A \subseteq B$.

For a supermodular similarity, the corresponding distance is submodular

supermodular $\not\Rightarrow$ metric (Berman & Blaschko, arXiv:1807.06686)

[Yu & Blaschko, ICML 2015; PAMI 2018]

Submodular Hamming distance

Definition (Submodular Hamming distance (Gillenwater et al., 2015))

Given a positive, monotone submodular set function g s.t. $g(\emptyset) = 0$, the corresponding submodular Hamming distance is $d_g(X, Y) := g(X \triangle Y)$.

Definition (Supermodular Hamming similarity)

A similarity S is called a supermodular Hamming similarity if $S(X, Y) = 1 - d_g(X, Y)$ for some submodular Hamming distance d_g.

Supermodular Hamming similarity

Theorem (Gillenwater et al., 2015)

For a supermodular Hamming similarity S, $1 - S$ is a (pseudo)metric.

Proof.

Denote $f = 1 - g$.

$$1 - S(X, Z) \leq 1 - S(X, Y) + 1 - S(Y, Z) \implies \tag{1}$$
$$f(X \triangle Y) + f(Y \triangle Z) \leq f(X \triangle Z) + 1. \tag{2}$$

Generalization of triangle inequality: $X \triangle Z \subseteq (X \triangle Y) \cup (Y \triangle Z)$
monotonicity of f: $f(X \triangle Z) \geq f((X \triangle Y) \cup (Y \triangle Z))$.
supermodularity of f:

$$f(X \triangle Y) + f(Y \triangle Z) \leq \underbrace{f((X \triangle Y) \cup (Y \triangle Z))}_{\leq f(X \triangle Z)} + \underbrace{f((X \triangle Y) \cap (Y \triangle Z))}_{\leq 1}$$

$\square$

Rational set similarities

name	$S(X,Y)$ $(X \neq Y)$	Submodularity w.r.t. $X \triangle Y$	LSHable
Jaccard	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\lvert X\triangle Y\rvert}$	Supermodular [22, Proposition 11]	yes
Hamming	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+\lvert X\triangle Y\rvert}$	Modular (Section 3)	yes
Anderberg	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+2\lvert X\triangle Y\rvert}$	Supermodular (Proposition 1)	yes
Rogers–Tanimoto	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+2\lvert X\triangle Y\rvert}$	Supermodular (Corollary 1)	yes
Simpson	$\frac{\lvert X\cap Y\rvert}{\min(\lvert X\rvert,\lvert Y\rvert)}$	Neither submodular nor supermodular (Proposition 3)	no
Braun–Blanquet	$\frac{\lvert X\cap Y\rvert}{\max(\lvert X\rvert,\lvert Y\rvert)}$	Neither submodular nor supermodular (Proposition 4)	no
Sørensen-Dice	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\frac{1}{2}\lvert X\triangle Y\rvert}$	Neither submodular nor supermodular [23, Proposition 6]	no
Sokal–Sneath 1	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+\frac{1}{2}\lvert X\triangle Y\rvert}$	Submodular (Corollary 2)	no
Forbes	$\frac{\lvert V\rvert\cdot\lvert X\cap Y\rvert}{\lvert X\rvert\cdot\lvert Y\rvert}$	Neither submodular nor supermodular (Proposition 5)	no
Sørensen$_\gamma$	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\gamma\lvert X\triangle Y\rvert}$	Supermodular for $\gamma \geq 1$, neither submodular nor supermodular for $0 < \gamma < 1$ (Proposition 1)	iff $\gamma \geq 1$
Sokal–Sneath$_\gamma$	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+\gamma\lvert X\triangle Y\rvert}$	Supermodular for $\gamma \geq 1$, Submoduar for $0 < \gamma < 1$ (Proposition 2)	iff $\gamma \geq 1$
Cardinality Intersection	Definition 5	Neither submodular nor supermodular (Proposition 6)	yes
Identity Intersection	Definition 6	Supermodular (Proposition 7)	yes

Berman, M. and M. B. Blaschko, arXiv:1807.06686; F. Chierichetti, R. Kumar, A. Panconesi, and E. Terolli, 2017

LSH preserving functions

Definition (LSH-preserving function)

A function $f : [0,1) \to [0,1]$ is LSH-preserving if $f \circ S$ is LSHable whenever S is LSHable.

Definition (Probability generating function)

A function $f(x)$ is a probability generating function (PGF) if there is a probabilty distribution $\{p_i\}_{0 \le i < \infty}$ such that $f(x) = \sum_{i=0}^{\infty} p_i x^i$ for $x \in [0,1]$.

Theorem (Theorem 3.1, Chierichetti & Kumar, 2012)

A function $f : [0,1) \to [0,1]$ is LSH-preserving iff there are a PGF p and a scalar $\alpha \in [0,1]$ such that $f(x) = \alpha p(x)$.

LSH-preserving functions are supermodular-preserving functions

Proposition (LSH-preserving functions are supermodularity-preserving functions)

Given an LSH-preserving function $f : [0, 1) \to [0, 1]$ and a non-negative monotonically decreasing supermodular function g such that $g(\emptyset) = 1$, $f \circ g$ is a non-negative monotonically decreasing supermodular function with $f \circ g(A) \in [0, 1]$ for all $A \subseteq V$.

Berman & Blaschko, arXiv:1807.06686

LSHability and supermodularity

Supermodularity $\not\Longrightarrow$ metric

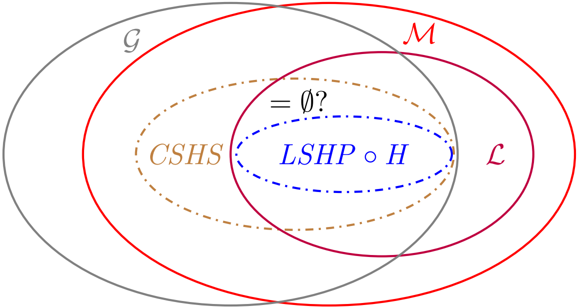
LSHable $\Longrightarrow$ metric

LSH-preserving = supermodular-preserving

LSHability and supermodularity 1-to-1 in the table of popular similarities

Metric supermodular $\Longleftrightarrow$ LSHable?

Our universe of similarities



Berman, M. and M. B. Blaschko: arXiv:1807.06686.

Proof technique - LSHability

Definition (Complete hash)

For a fixed $d = |\mathcal{X}|$, we define a complete hash as a set of hash functions $\mathcal{H}$ such that for all partitions of $\mathcal{X}$, there exists $h \in \mathcal{H}$ such that $h(x_i) = h(x_j)$ iff $x_i, x_j \in \mathcal{X}$ are in the same subset of the partition.

The size of $\mathcal{H}_d$ is given by the dth Bell number, which satisfies the recurrence $B_0 = 1$,

$$B_d = \sum_{k=0}^{d-1} \binom{d-1}{k} B_k. \tag{3}$$

Exponential in d.

Complete hash: example for $|\mathcal{X}| = 4$

$$h_1(\emptyset) = 1,\ h_1(\{1\}) = 1,\ h_1(\{2\}) = 1,\ h_1(\{1,2\}) = 1;$$
$$h_2(\emptyset) = 1,\ h_2(\{1\}) = 1,\ h_2(\{2\}) = 1,\ h_2(\{1,2\}) = 2;$$
$$h_3(\emptyset) = 1,\ h_3(\{1\}) = 1,\ h_3(\{2\}) = 2,\ h_3(\{1,2\}) = 1;$$
$$h_4(\emptyset) = 1,\ h_4(\{1\}) = 1,\ h_4(\{2\}) = 2,\ h_4(\{1,2\}) = 2;$$
$$h_5(\emptyset) = 1,\ h_5(\{1\}) = 1,\ h_5(\{2\}) = 2,\ h_5(\{1,2\}) = 3;$$
$$h_6(\emptyset) = 1,\ h_6(\{1\}) = 2,\ h_6(\{2\}) = 1,\ h_6(\{1,2\}) = 1;$$
$$h_7(\emptyset) = 1,\ h_7(\{1\}) = 2,\ h_7(\{2\}) = 1,\ h_7(\{1,2\}) = 2;$$
$$h_8(\emptyset) = 1,\ h_8(\{1\}) = 2,\ h_8(\{2\}) = 1,\ h_8(\{1,2\}) = 3;$$
$$h_9(\emptyset) = 1,\ h_9(\{1\}) = 2,\ h_9(\{2\}) = 2,\ h_9(\{1,2\}) = 1;$$
$$h_{10}(\emptyset) = 1,\ h_{10}(\{1\}) = 2,\ h_{10}(\{2\}) = 2,\ h_{10}(\{1,2\}) = 2;$$
$$h_{11}(\emptyset) = 1,\ h_{11}(\{1\}) = 2,\ h_{11}(\{2\}) = 2,\ h_{11}(\{1,2\}) = 3;$$
$$h_{12}(\emptyset) = 1,\ h_{12}(\{1\}) = 2,\ h_{12}(\{2\}) = 3,\ h_{12}(\{1,2\}) = 1;$$
$$h_{13}(\emptyset) = 1,\ h_{13}(\{1\}) = 2,\ h_{13}(\{2\}) = 3,\ h_{13}(\{1,2\}) = 2;$$
$$h_{14}(\emptyset) = 1,\ h_{14}(\{1\}) = 2,\ h_{14}(\{2\}) = 3,\ h_{14}(\{1,2\}) = 3;$$
$$h_{15}(\emptyset) = 1,\ h_{15}(\{1\}) = 2,\ h_{15}(\{2\}) = 3,\ h_{15}(\{1,2\}) = 4,$$

Proof technique - LSHability

$A \in \mathbb{R}^{\binom{d}{2} \times B_d}$:

$$A_{(i,j),k} = \begin{cases} 1 & \text{if } H_{ik} = H_{jk}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$b \in \mathbb{R}^{\binom{d}{2}}$:

$$b_{(i,j)} = S(i,j). \tag{5}$$

Proposition

A similarity $S : \mathcal{X} \times \mathcal{X} \to [0,1]$ is LSHable iff for A and b defined as in Equations (4) and (5), the following linear system is feasible for some $x \in \mathbb{R}^{B_d}$:

$$\forall i,\ x_i \geq 0, \quad \sum_{i=1}^{B_d} x_i = 1, \quad Ax = b. \tag{6}$$

Furthermore, for any x satisfying this linear system, $P_{\mathcal{H}}(h) = x_h$ is a valid LSH for S.

Proof technique

- Properties characterized by an (exponential sized) set of linear constraints on the similarity matrix
- Exhaustive search over a good guess of potential counterexamples

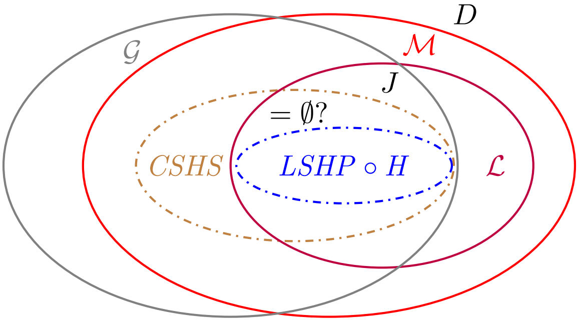
Proposition (Berman & Blaschko, 2018)

That a similarity is metric supermodular does not imply that it is LSHable.

Proof.

We prove this with a counterexample that is metric supermodular but not LSHable: $S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \gamma \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \gamma \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1-\gamma \\ 0 & 0 & 0 & \gamma & 0 & \gamma & 1-\gamma & 1 \end{pmatrix}$, where e.g. $\gamma = 1/8$. $\square$

Jaccard and Dice



Berman & Blaschko, arXiv:1807.06686; Yu & Blaschko, ICML 2015; AISTATS 2016; PAMI 2018.

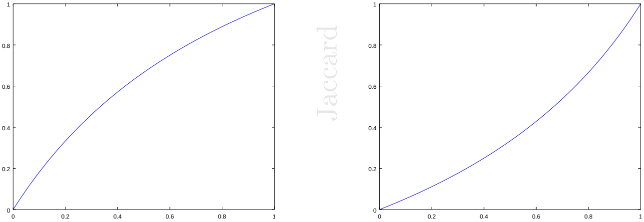
Relationship between Jaccard and Dice

$$D(y,\tilde{y}) := \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|},\ J(y,\tilde{y}) := \frac{|y \cap \tilde{y}|}{|y \cup \tilde{y}|},\ H(y,\tilde{y}) := 1 - \frac{|y \setminus \tilde{y}| + |\tilde{y} \setminus y|}{d}, \tag{7}$$

$$H_\gamma(y,\tilde{y}) := 1 - \gamma\frac{|y \setminus \tilde{y}|}{|y|} - (1-\gamma)\frac{|\tilde{y} \setminus y|}{d - |y|}, \tag{8}$$

$D(y,\tilde{y}) = \frac{2J(y,\tilde{y})}{1+J(y,\tilde{y})}$ and $J(y,\tilde{y}) = \frac{D(y,\tilde{y})}{2-D(y,\tilde{y})}$

Dice



Jaccard

Jaccard

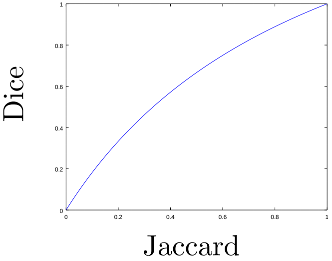
Dice

Relationship between Jaccard and Dice

$$D(y,\tilde{y}) := \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|},\ J(y,\tilde{y}) := \frac{|y \cap \tilde{y}|}{|y \cup \tilde{y}|},\ H(y,\tilde{y}) := 1 - \frac{|y \setminus \tilde{y}| + |\tilde{y} \setminus y|}{d}, \tag{7}$$

$$H_\gamma(y,\tilde{y}) := 1 - \gamma \frac{|y \setminus \tilde{y}|}{|y|} - (1-\gamma)\frac{|\tilde{y} \setminus y|}{d - |y|}, \tag{8}$$

$$D(y,\tilde{y}) = \frac{2J(y,\tilde{y})}{1+J(y,\tilde{y})} \text{ and } J(y,\tilde{y}) = \frac{D(y,\tilde{y})}{2-D(y,\tilde{y})}$$



Jaccard and Dice - approximation

Definition (Absolute approximation)

A similarity S is absolutely approximated by $\tilde{S}$ with error $\varepsilon \geq 0$ if the following holds for all y and $\tilde{y}$:

$$|S(y, \tilde{y}) - \tilde{S}(y, \tilde{y})| \leq \varepsilon. \tag{9}$$

Definition (Relative approximation)

A similarity S is relatively approximated by $\tilde{S}$ with error $\varepsilon \geq 0$ if the following holds for all y and $\tilde{y}$:

$$\frac{\tilde{S}(y, \tilde{y})}{1 + \varepsilon} \leq S(y, \tilde{y}) \leq \tilde{S}(y, \tilde{y})(1 + \varepsilon). \tag{10}$$

Proposition

J and D approximate each other with relative error of 1 and absolute error of $3 - 2\sqrt{2} = 0.17157\ldots$.

Jaccard, Dice, and weighted-Hamming

Defining "distortion" of an approximation as a one-sided version of our definition of a relative approximation:

Theorem (Chierichetti et al., 2017)

Jaccard is the minimum-distortion LSHable approximation to Dice

Proposition

D and H_γ (where γ is chosen to minimize the approximation factor between D and H_γ) do not relatively approximate each other, and absolutely approximate each other with an error of 1. We note that the absolute error bound is trivial as D and H_γ are both similarities in the range $[0, 1]$.

Jaccard, Dice, and weighted-Hamming

Defining “distortion” of an approximation as a one-sided version of our definition of a relative approximation:

Theorem (Chierichetti et al., 2017)

Jaccard is the minimum-distortion LSHable approximation to Dice

Proposition

D and H_γ (where γ is chosen to minimize the approximation factor between D and H_γ) do not relatively approximate each other, and absolutely approximate each other with an error of 1. We note that the absolute error bound is trivial as D and H_γ are both similarities in the range $[0, 1]$.

Regularized risk

Consider a population distribution $P(x, y)$ and an empirical measure from a sample of size n, $P_n(x, y)$.

Definition (Risk)

For a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, the population (true) risk of a function $f : \mathcal{X} \to \mathcal{Y}$ is

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P}\left[\Delta(f(x), y)\right] \tag{11}$$

We may similarly consider the *empirical* risk

$$\hat{\mathcal{R}}(f) := \mathbb{E}_{(x,y)\sim P_n}\left[\Delta(f(x), y)\right] \tag{12}$$

In practice, we optimize something like

$$\arg\min_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P_n}\left[\ell(f(x), y)\right] + \lambda\Omega(f) \tag{13}$$

where $\lambda > 0$ is chosen by a model selection procedure, and ℓ is a tractable (at least differentiable a.e. and not piecewise constant) surrogate to Δ.

Regularized risk

Consider a population distribution $P(x, y)$ and an empirical measure from a sample of size n, $P_n(x, y)$.

Definition (Risk)

For a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, the population (true) risk of a function $f : \mathcal{X} \to \mathcal{Y}$ is

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P} \left[\Delta(f(x), y)\right] \tag{11}$$

We may similarly consider the *empirical* risk

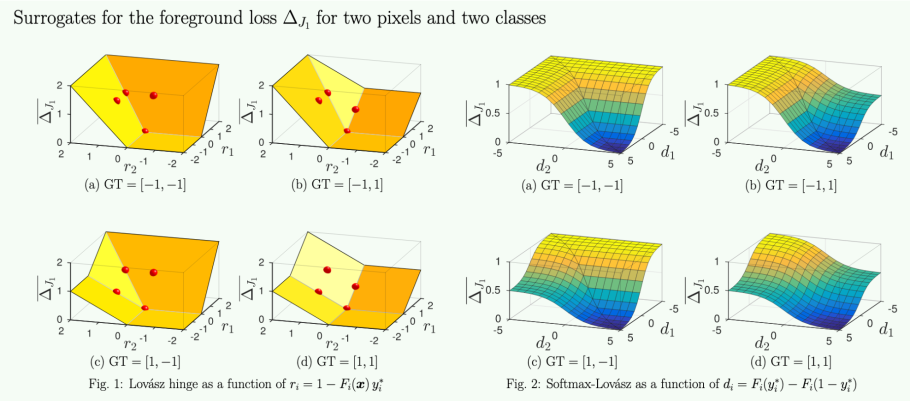
$$\hat{\mathcal{R}}(f) := \mathbb{E}_{(x,y)\sim P_n} \left[\Delta(f(x), y)\right] \tag{12}$$

In practice, we optimize something like

$$\arg\min_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P_n} \left[\ell(f(x), y)\right] + \lambda\Omega(f) \tag{13}$$

where $\lambda > 0$ is chosen by a model selection procedure, and ℓ is a tractable (at least differentiable a.e. and not piecewise constant) surrogate to Δ.

Lovász hinge and Lovász-Softmax



[Yu & Blaschko 2015; 2018; Berman, Rannen Triki, & Blaschko CVPR 2018]

Multi-class extension

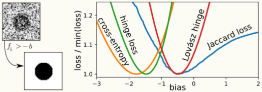
$$M_c(y, \tilde{y}) = \{y = c, \tilde{y} \neq c\} \cup \{y \neq c, \tilde{y} = c\}$$

$$\Delta_J(y, \tilde{y}) = \sum_{j=1}^{k} \frac{|M_j(y, \tilde{y})|}{|\{y = c\} \cup M_j(y, \tilde{y})|}$$

[Berman et al., CVPR 2018]

Jaccard results

7. Binary toy experiment

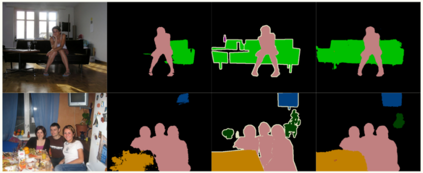


8. Pascal VOC binary experiment

Training loss →	Cross-entropy	Hinge	Lovász hinge
Cross-entropy	**6.8**	7.0	8.0
Hinge	7.8	**7.0**	7.1
Lovász hinge	8.4	7.5	**5.4**
Image–IoU (%)	77.1	75.8	**80.5**

9. Pascal VOC multiclass exp.

- Network: DeepLab-v2 single-scale [2])



(a) image (b) Cross-entropy (c) ground truth (d) Lovász-Softmax

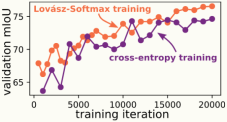


Fig. 3: Validation mIoU evolution

- Pascal VOC test server mIoU increased from 76.4% to 79.0%

What about Dice?

Jaccard has many favorable properties, but medical legacy of Dice won't be wiped away overnight

Optimizing Jaccard minimizes an upper bound on Dice:

$$1 - D(y, \tilde{y}) \leq 1 - J(y, \tilde{y}) \implies$$
$$\mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))] \leq \mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))]$$

Optimizing Dice minimizes an upper bound on Jaccard:

$$\varphi(x) = 2x/(1+x)$$

Jensen's inequality:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))] &= \mathbb{E}_{(x,y)\sim P_n}[\varphi(1 - D(y, f(x)))] \\ &\leq \varphi(\mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))])\end{aligned}$$

φ monotonic over $[0, 1] \implies$ for every λ in $\min_f \varphi(\hat{\mathcal{R}}(f)) + \lambda\Omega(f)$ there exists $\tilde{\lambda}$ s.t. $\min_f \hat{\mathcal{R}}(f) + \tilde{\lambda}\Omega(f)$ has the same minimizer

What about Dice?

Jaccard has many favorable properties, but medical legacy of Dice won't be wiped away overnight

Optimizing Jaccard minimizes an upper bound on Dice:

$$\begin{aligned} &1 - D(y, \tilde{y}) \leq 1 - J(y, \tilde{y}) \implies \\ &\quad \mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))] \leq \mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))] \end{aligned}$$

Optimizing Dice minimizes an upper bound on Jaccard:

$$\varphi(x) = 2x/(1+x)$$

Jensen's inequality:

$$\begin{aligned} \mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))] &= \mathbb{E}_{(x,y)\sim P_n}[\varphi(1 - D(y, f(x)))] \\ &\leq \varphi(\mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))]) \end{aligned}$$

φ monotonic over $[0,1] \implies$ for every λ in $\min_f \varphi(\hat{\mathcal{R}}(f)) + \lambda\Omega(f)$ there exists $\tilde{\lambda}$ s.t. $\min_f \hat{\mathcal{R}}(f) + \tilde{\lambda}\Omega(f)$ has the same minimizer

What about Dice?

Jaccard has many favorable properties, but medical legacy of Dice won't be wiped away overnight

Optimizing Jaccard minimizes an upper bound on Dice:

$$
\begin{aligned}
&1 - D(y, \tilde{y}) \leq 1 - J(y, \tilde{y}) \implies \\
&\quad \mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))] \leq \mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))]
\end{aligned}
$$

Optimizing Dice minimizes an upper bound on Jaccard:

$$\varphi(x) = 2x/(1+x)$$

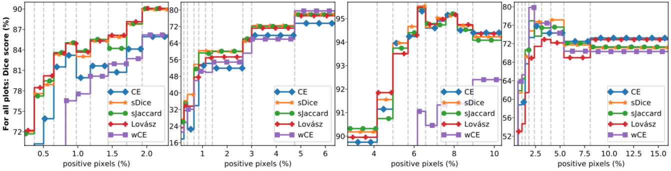
Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))] &= \mathbb{E}_{(x,y)\sim P_n}[\varphi(1 - D(y, f(x)))] \\
&\leq \varphi(\mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))])
\end{aligned}
$$

φ monotonic over $[0,1] \implies$ for every λ in $\min_f \varphi(\hat{\mathcal{R}}(f)) + \lambda\Omega(f)$ there exists $\tilde{\lambda}$ s.t. $\min_f \hat{\mathcal{R}}(f) + \tilde{\lambda}\Omega(f)$ has the same minimizer

Dice results

	Dataset	*loss* → CE	wCE	sDice	sJaccard	Lovász
Dice score	BR18	*0.768*	*0.735*	0.823	0.823	**0.827**
	IS17	*0.260*	0.311	0.331	0.321	0.305
	IS18	*0.463*	*0.474*	**0.538**	0.528	0.508
	MO17	0.930	*0.860*	0.932	0.931	0.932
	PO18	*0.635*	*0.602*	0.656	0.651	0.649
Jaccard index	BR18	*0.654*	*0.602*	0.717	0.720	0.722
	IS17	*0.177*	0.212	0.227	0.217	0.204
	IS18	*0.345*	*0.344*	**0.407**	0.399	0.382
	MO17	0.873	*0.769*	0.877	0.875	0.877
	PO18	*0.541*	*0.488*	0.559	0.554	0.553



(a) BRATS 2018 (b) ISLES 2018 (c) MO17 (d) PO18

77 learning-based segmentation papers in MICCAI 2018 - evaluate with Dice
47 trained using per-pixel loss

[Bertels et al., under review 2019]

Lovász-Softmax code - PyTorch & TensorFlow
`https://github.com/bermanmaxim/LovaszSoftmax`

We're looking for grad students to start as early as Oct, 2019
Apply directly by emailing a CV

Matthew Blaschko
`http://homes.esat.kuleuven.be/~mblaschk/`
matthew.blaschko@esat.kuleuven.be