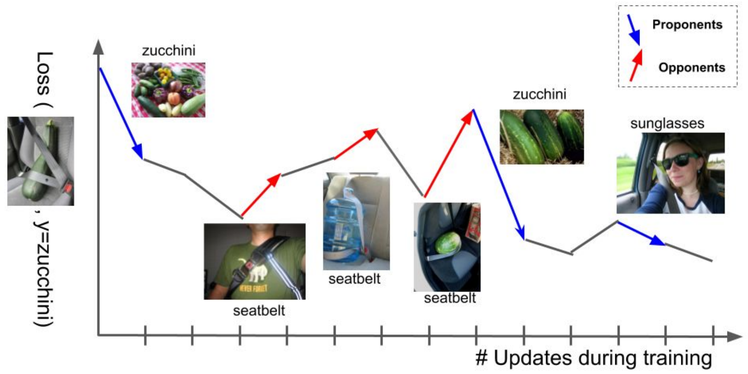
# TracIn: Estimating Training Data Influence by Tracing Gradient Descent

Garima Pruthi, Frederick Liu, Satyen Kale, Mukund Sundararajan
Contact: mukunds@google.com

NEURAL INFORMATION PROCESSING SYSTEMS

## How does a training point influence a prediction?

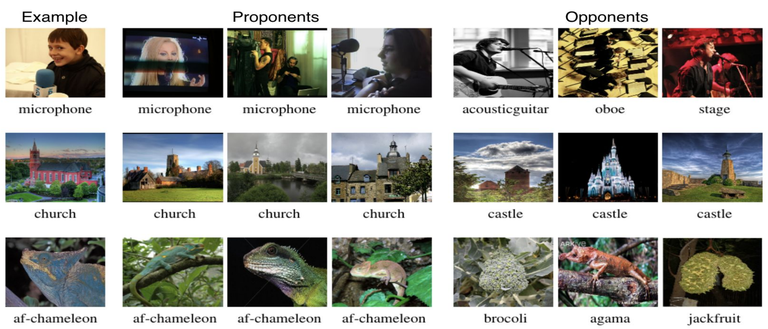
**Influence of a training point on a prediction is the change in loss of prediction point when visited during gradient descent.**




## Proponents and Opponents

Proponents reduce loss ⇒ positive influence score
Opponents enlarge loss ⇒ negative influence score

**ImageNet**




**DBPedia**

| | | |
|---|---|---|
| Example | OfficeHolder | **Manuel Azaña** Manuel Azaña Díaz (Alcalá de Henares January 10 1880 – Montauban November 3 1940) was the first Prime Minister of the Second Spanish Republic (1931–1933) and later served again as Prime Minister (1936) and then as the second and last President of the Republic (1936–1939). The Spanish Civil War broke out while he was President. With the defeat of the Republic in 1939 he fled to France resigned his office and died in exile shortly afterwards. |
| Opponents | Artist | **Mikołaj Rej** Mikołaj Rej or Mikołaj Rey of Nagłowice (February 4 1505 – between September 8 and October 5 1569) was a Polish poet and prose writer of the emerging Renaissance in Poland as it succeeded the Middle Ages as well as a politician and musician. He was the first Polish author to write exclusively in the Polish language and is considered (with Biernat of Lublin and Jan Kochanowski) to be one of the founders of Polish literary language and literature. |
| Opponents | Artist | **Justin Jeffre** Justin Paul Jeffre (born on February 25 1973) is an American pop singer and politician. A long-time resident and vocal supporter of Cincinnati Jeffre is probably best known as a member of the multi-platinum selling boy band 98 Degrees.Before shooting to super stardom Jeffre was a student at the School for Creative and Performing Arts in Cincinnati. It was there that he first became friends with Nick Lachey. The two would later team up with Drew Lachey and Jeff Timmons to form 98 Degrees. |
| Opponents | Artist | **David Kitt** David Kitt (born 1975 in Dublin Ireland) is an Irish musician. He is the son of Irish politician Tom Kitt.He has released six studio albums to date: Small Moments The Big Romance Square 1 The Black and Red Notebook Not Fade Away and The Nightsaver. |

## TracIn Influence score

training point — summing over time step when training point z is used

$$\texttt{TracInIdeal}(z, z') = \sum_{t:\ z_t = z} \ell(w_t, z') - \ell(w_{t+1}, z')$$

test point — loss of test point at time step t — loss of test point at time step t+1

## A Scalable, Simple, and General implementation

- **Scalable → Leveraging gradients**

**First Order Approximation**

$$\ell(w_{t+1}, z') = \ell(w_t, z') + \nabla\ell(w_t, z') \cdot (w_{t+1} - w_t) + O(\|w_{t+1} - w_t\|^2)$$

**Stochastic Gradient Descent**

$$w_{t+1} - w_t = -\eta_t \nabla\ell(w_t, z_t)$$

**Approximate TracInIdeal**

$$\ell(w_t, z') - \ell(w_{t+1}, z') \approx \eta_t \nabla\ell(w_t, z') \cdot \nabla\ell(w_t, z_t)$$

$$\texttt{TracIn}(z, z') = \sum_{t:\ z_t = z} \eta_t \nabla\ell(w_t, z') \cdot \nabla\ell(w_t, z)$$

- **Simple**
  → **Leveraging checkpoints**
  → **No Hessian (Influence functions, Koh et al, ICML 2017)**
  → **No Optimality Conditions (Representer Points, Yeh et al, NeurIPS 2018)**

**Practical Heuristic Influence via Checkpoints - TracInCP**

test point — learning rate at checkpoint i

$$\texttt{TracInCP}(z, z') = \sum_{i=1}^{k} \eta_i \nabla\ell(w_{t_i}, z) \cdot \nabla\ell(w_{t_i}, z')$$

training point — loss gradient for test point at checkpoint i — loss gradient for training point at checkpoint i — summing over checkpoints

- **General**
  → **Works for any model trained with Gradient Descent-like optimizer.**
  → **Influence over any differentiable metric can be computed.**

## Extending TracIn with Self-Influence

- **Influence of training point on itself (z=z')**
- **Memorization manifests as high self-influence**
- **Memorization implies rarity in feature, label space**

**High Self-Influence examples from ImageNet**

**Multiple classes of objects**

Label: tennis ball
Pred: racket



Label: neck brace
Pred: crutch



Label: zucchini
Pred: seatbelt

**Whole vs. part**

Label: grocery store
Pred: crab

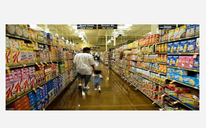
Label: barber shop
Pred: barber chair

Label: knee pad
Pred: volleyball

**Mislabelled**

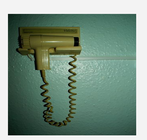
Label: barbell
Pred: dumbell

Label: hotpot
Pred: crockpot



Label: confectionery
Pred: grocery store

**Unique within label class**



Label: hand blower
Pred: hand blower

Label: palace
Pred: palace

Label: oscilloscope
Pred: oscilloscope

TracIn is a **scalable, simple,** and **general** method to estimate influence.

Code at: https://github.com/frederick0329/TracIn