Information k-means, fragmentation and syntax analysis: A new approach to unsupervised Machine Learning

Gautier Appert

PhD supervisor : Olivier Catoni

October 29, 2020

Purpose

Given a sample $X_1, \ldots, X_n$

of n independent copies of a signal $X \in \mathbb{R}^d$ (we will focus on digital images), we want to produce a better representation of its content.
We want to

1. identify significant patterns
2. represent their interactions

We propose to

1. perform some vector quantization of image fragments
2. compute syntax trees for the quantized fragments

both steps being based on a data compression criterion (performing lossy coding in the first step and lossless coding in the second one).

In the large sample limit we can recover the signal distribution

- Let $\overline{X} = (X_1, \dots, X_n) \in \mathbb{R}^{nd}$ and let $\theta(\overline{X})$ be a (lossy) binary prefix code for $\overline{X}$ (we code the whole sample $\overline{X}$, not a single random image X).
- Let $\overline{Y} = f\left[\theta(\overline{X})\right]$ be the lossy decoding of $\overline{X}$ from its binary representation $\theta(\overline{X})$.
- Assume that $\left\|\overline{Y} - \overline{X}\right\|^2 \leq nd\alpha$, for some distortion level $\alpha > 0$.
- Introduce a coding distribution Q_θ. Assume that $Q_{\overline{Y}} = Q_\theta \circ f^{-1} = Q_{f(\theta)}$ is exchangeable.
- Consider a blurred version of the sample $\overline{X}$ defined as $\overline{V} = (V_1, \dots, V_n) = \overline{X} + \overline{W}$, where $\overline{W}$ is independent of the sample $\overline{X}$ and distributed as $\mathcal{N}\left(0, \sigma^2 I_{nd}\right)$.
- Define $Q_{\overline{V}}$ from $Q_{\overline{Y}}$, setting $Q_{\overline{V}|\overline{Y}} = \mathcal{N}\left(\overline{Y}, \sigma^2 I_{nd}\right)$.

Lemma : The progressive estimator

$$\overline{Q}_{V_n \mid V_1,\dots,V_{n-1}} = \frac{1}{n}\sum_{i=1}^{n} Q_{V_n \mid V_1,\dots,V_{i-1}}$$

is such that

$$\frac{1}{d}\mathbb{P}_{V_1,\dots,V_{n-1}}\left[\mathcal{K}\left(\mathbb{P}_{V_n}, \overline{Q}_{V_n \mid V_1,\dots,V_{n-1}}\right)\right] \leq \frac{1}{nd}\mathcal{K}\left(\mathbb{P}_\theta, Q_\theta\right) + \frac{\alpha}{2\sigma^2}.$$

The estimator $\overline{Q}$ is successful at approaching $\mathbb{P}_V$ when Q_θ is a successful data compression scheme for the source $\mathbb{P}_\theta = \mathbb{P}_{\theta(\overline{X})}$. Moreover $\overline{Q}$ is a function of Q_θ, so that Q_θ contains in this case enough information to characterize $\mathbb{P}_V$ that is itself an approximation of $\mathbb{P}_X$.

Explanations: statistical estimation without statistical modeling

- $\left\{ \text{ statistical models } \right\} \subsetneq \left\{ \text{ compression models } \right\}$ . . . and both can produce estimators, according to the previous lemma.
- Namely, if we use a parametric model $\{Q_{\theta\,|\,\beta}\}$, hoping that $\inf_\beta \mathcal{K}\big(\mathbb{P}_\theta, Q_{\theta\,|\,\beta}\big)$ is small and if we set $Q_\theta = \int Q_{\theta\,|\,\beta}\,\mathrm{d}\mu(\beta)$ for some prior probability measure μ on the parameter β, the lemma produces a statistical estimator $\overline{Q}$ from a statistical model $\{Q_{\theta\,|\,\beta}\}$.
- But we can choose Q_θ otherwise, and in particular we can use a grammar based code. The general principle is to define small binary indexes for repeated large patterns and to gather those possibly nested indexation rules into a grammar. A widely used grammar based code is the Lempel Ziv algorithm. We will propose a more evolved choice of grammar producing syntax trees related to the conditional probability of observing some selected patterns in a given context.

Expected benefits

1. Grammar based compression can process in a meaningful way infrequent patterns : even seen in the sample only twice, a large pattern is worth being indexed by a small index and described only once. On the other hand a statistical estimator, based on the estimation of expectations, cannot take into account in a reliable way an event that occurs only twice in the sample.
2. A parametric model $Q_{\theta \mid \beta}$ that accounts for elaborate conditional independence assumptions will presumably depend on a high-dimensional parameter β, and estimation in high dimension requires huge samples unless the complexity is restricted in some other way.

For those two reasons, we hope to make sense of small samples compared to a more traditional statistical approach.

First lossy compression step : fragmentation

- Let $\mathcal{C} \subset \mathbb{R}$ be finite (or more generally countable). For instance $\mathcal{C} = [\![0, 255]\!]$.
- We code the sample $\overline{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{nd}$ by

$$\theta = \Big[\big(A_i \subset [\![1,k]\!], 1 \leq i \leq n\big), \big(B_j \subset [\![1,d]\!], C_j \in \mathcal{C}^{B_j}, 1 \leq j \leq k\big)\Big],$$

where $A_i \in \mathcal{T}_B = \left\{A \subset [\![1,k]\!] : \bigsqcup_{j \in A} B_j = [\![1,d]\!]\right\}$,

- the lossy decoding function being

$$f(\theta) = \overline{Y} = \Big(C_{A_i}, 1 \leq i \leq n\Big), \text{ where } C_A = \sum_{j \in A} C_j,$$

and where C_j is set to 0 outside its support B_j.

Distortion and k-means

- The distortion $\mathcal{D}(\overline{X}, \theta) = (nd)^{-1}\|\overline{X} - \overline{Y}\|^2$ is minimized for a given set of fragments (B, C) when $A_i \in \arg\min_{A \in \mathcal{T}_B} \|X_i - C_A\|$ and is then equal to

$$\mathcal{D}(\overline{X}, B, C) = \inf_{A \in \mathcal{T}_B^n} \mathcal{D}\big(\overline{X}, \underbrace{(A, B, C)}_{\theta}\big) = d^{-1}\overline{\mathbb{P}}_X\Big(\min_{A \in \mathcal{T}_B} \big\|X - C_A\big\|^2\Big),$$

  where $\overline{\mathbb{P}}_X = \dfrac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure.
- We see that the distortion is a generalization of the k-means empirical criterion, that we get when $B_j = [\![1, d]\!]$, $j \in [\![1, k]\!]$.

Principle of a fragmentation algorithm

- Create a sequence of codes $\theta_k = (A_k, B_k, C_k), k \geq n$.
- Start with one fragment per image: $k = n$, $A_{n,i} = \{i\}$, $i \in [\![1, n]\!]$, $B_{n,j} = [\![1, d]\!]$, $C_{n,j,s} = \arg\min_{c \in \mathcal{C}} |X_{j,s} - c|$, $j \in [\![1, k]\!]$.
- Maintain through iterations on k the two properties
  - $\mathcal{D}(\overline{X}, \theta_k) \leq \alpha$
  - and $[\![1, d]\!] = \bigsqcup_{j \in A_{k,i}} B_{k,j}$, $i \in [\![1, n]\!]$,

  while minimizing the code length of θ_k at each step.
- Stop when it is no more possible to decrease the code length of θ_k or earlier for a partial fragmentation.

The fragmentation loop

Iterate on $k \geq n$ the following.

- Choose a pair $J_k \subset [\![1, k]\!]$,
- set $A_{k+1,i} = A_{k,i} \cup \{k+1\}$ if $A_{k,i} \cap J_k \neq \varnothing$, $A_{k+1,i} = A_{k,i}$ otherwise,
- choose a support $B_{k+1,k+1} \subset \bigcap_{j \in J_k} B_{k,j}$,
- set $B_{k+1,j} = B_{k,j} \setminus B_{k+1,k+1}$ if $j \in J_k$, and $B_{k+1,j} = B_{k,j}$, if $j \in [\![1, k]\!] \setminus J_k$.
- set $C_{k+1,j,s} \in \arg\min_{c \in \mathcal{C}} \{ \left| c - \overline{\mathbb{P}}\left(X_{l,s} \,|\, j \in A_{k+1,l}\right) \right| \}$,

Choice of $B_{k+1,k+1}$, ensuring that $\mathcal{D}(\overline{X}, \theta_{k+1}) \leq \alpha$:

$$B_{k+1,k+1} = \Big\{ s \in \bigcap_{j \in J_k} B_{k,j} \,:\, \mathbf{Var}\big(\overline{\mathbb{P}}_{X_{l,s} \,|\, k+1 \in A_{k+1,l}}\big) + \min_{c \in \mathcal{C}} \big[c - \overline{\mathbb{P}}\big(X_{l,s} \,|\, k+1 \in A_{k+1,l}\big) \big]^2 \leq \alpha \Big\},$$

where $\overline{\mathbb{P}}_l = \dfrac{1}{n} \sum_{i=1}^{n} \delta_i$. $s \in B_j$.

Choice of J_k

- Let $|A_k| = \sum_{i=1}^{n} |A_{k,i}|$ and $|B_k| = \sum_{j=1}^{k} |B_{k,j}|$.
- Assuming that everything is coded on L bits, $Q(\theta_k) = 2^{-L\left(|A_k|+2|B_k|+(n+2k)\right)}$ is a sub-probability measure.
- Thus

$$\xi(k) = L^{-1} \log_2\left(Q(\theta_{k+1})/Q(\theta_k)\right)$$
$$= 2|B_{k+1,k+1}| - \sum_{i=1}^{n} \mathbb{1}\left(k+1 \in A_{k+1,i}\right) - 2$$

  can be maximized in J_k to find the pair giving the best data compression.
- A faster less optimal choice consists in choosing $J_k = \{j_{k,1}, j_{k,2}\}$, where $j_{k,1} \in \arg\max_j 2|B_{k,j}| - \sum_{i=1}^{n} \mathbb{1}\left(j \in A_{k,i}\right)$ and $j_{k,2} \in \arg\max_{j_{k,2}} \xi(k)$.

Purpose

We get from the first fragmentation step a lossy code $\theta = (A, B, C)$ made of a fragment codebook $(B, C) = (B_j, C_j)_{j=1}^{k}$ described by their supports B_j and their contents C_j, and n sets of labels $A = (A_i)_{i=1}^{n}$, where $A_i \subset [\![1, k]\!]$, describing the sample images. The aim of the syntax analysis step is to perform lossless data compression on the sample description A.

Starting point

The enumeration of A as a list of words with separators between the sets

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge .$$

This representation follows the syntax $\{\{w\}\wedge\}$ where $w \in [\![1, k]\!]$ is of type fragment label and where $\{\}$ denotes repetition, according to Extended Backus Naur specifications.

- Starting from $A_{0,i} = A_i$, $i \in [\![1, n]\!]$, iterate the following for $m \geq 0$.
  - Choose a pair $J_m \subset [\![1, k+m]\!]$ and define $A_{m+1,i} = \left(A_{m,i} \setminus J_m\right) \cup \{k+m+1\}$ when $J_m \subset A_{m,i}$ and $A_{m+1,i} = A_{m,i}$ otherwise.
  - Choose J_m to maximize $\sum_{i=1}^{n} \mathbb{1}\left(J_m \subset A_{m,i}\right)$ as long as it is greater than 2.
- Appending the description of the pairs, we get a representation of type $\{\{w|p\}\wedge\}\{ab\}$ where $w \in [\![1, k]\!]$, $p \in [\![k+1, k+m]\!]$ and $a, b \in [\![1, k+m]\!]$.
- Its length decreases at each step.

We appended to the sample representation $R ::= \{\{w|p\}\wedge\}$ containing non terminal symbols of type p, the description of a grammar $G ::= \{ab\}$ represented by the right-hand side of context free rewriting rules $p \to ab$.

Second step : grammar based grammar compression

- We can reindex the grammar G to get $a_1 b_{1,1} a_1 b_{1,2} \dots a_1 b_{1,q_1} \dots a_m b_{m,1} \dots a_m b_{m,q_m}$ where the a_i are distinct and $q_1, \dots, q_m$ are maximal.
- We can then factorize G into $a_1 \dots a_m \wedge b_{1,1} \dots b_{1,q_1} \wedge \dots \wedge b_{m,1} \dots b_{m,q_m} \wedge$ that is of the type $G ::= \{a\} \wedge \{\{b\} \wedge\}$.
- The second part of the description of G, describing the contexts $\{\{b\} \wedge\}$ of the pairs first elements $\{a\}$, is of the same type as the original representation and can be compressed again.
- Thus $\{\{b\} \wedge\}$ becomes $\{\{b|s\} \wedge\}\{cd\}$, where s are new non terminal syntax symbols and where $\{cd\}$ is a second grammar.
- The syntax labels s represent sets of configurations that appear in the same contexts $\{a\}$ and therefore perform a certain kind of context analysis.
- The whole representation follows the syntax $\{\{w|p\} \wedge\}\{a\} \wedge \{\{b|s\} \wedge\}\{cd\}$.

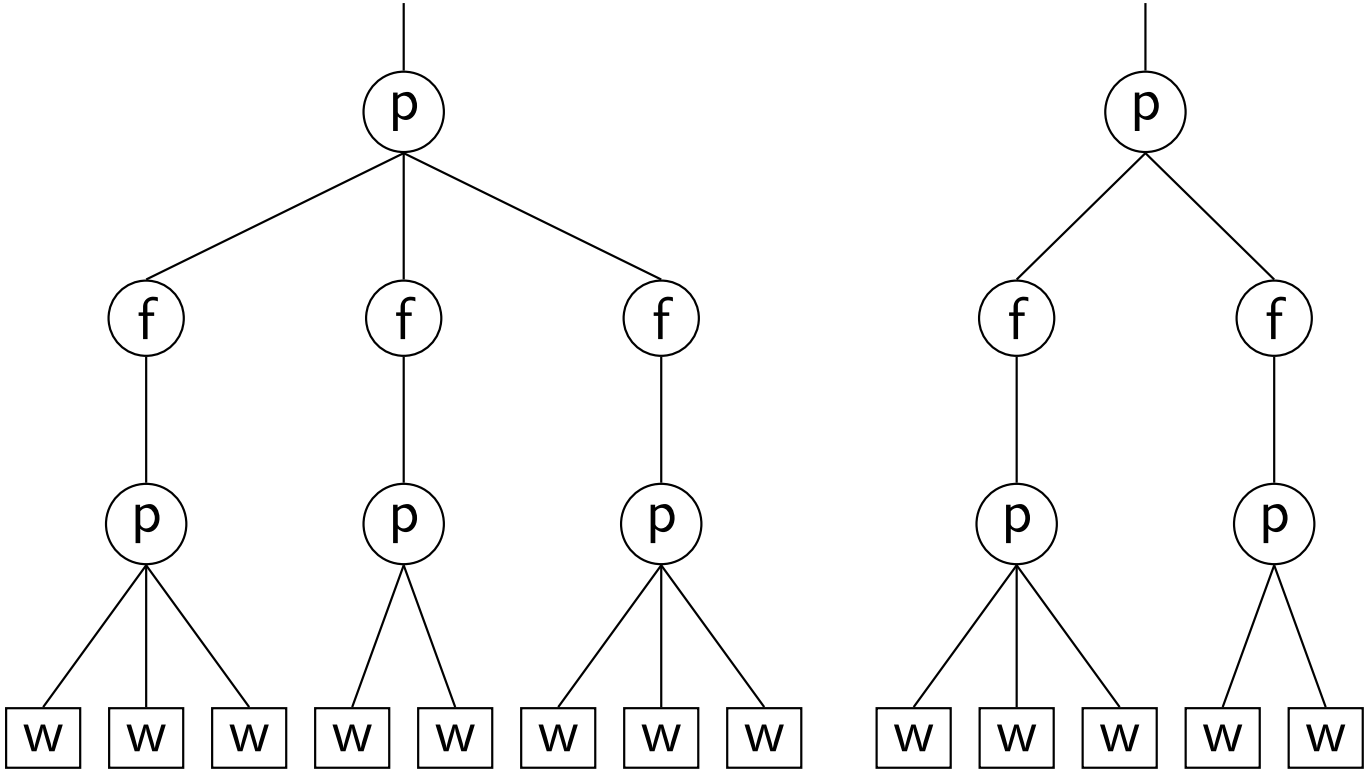
Syntax labels

The syntax labels of step 2 induce a classification of the symbols of step 1. We define

- $f(w) = w$ when $w \in [\![1, k]\!]$,
- $f(p) = s$ when $p \to ab$ and b is recovered from the compressed representation of ab by rewriting s
- and $f(p) = b$ when $p \to ab$ has not been compressed in the second step.

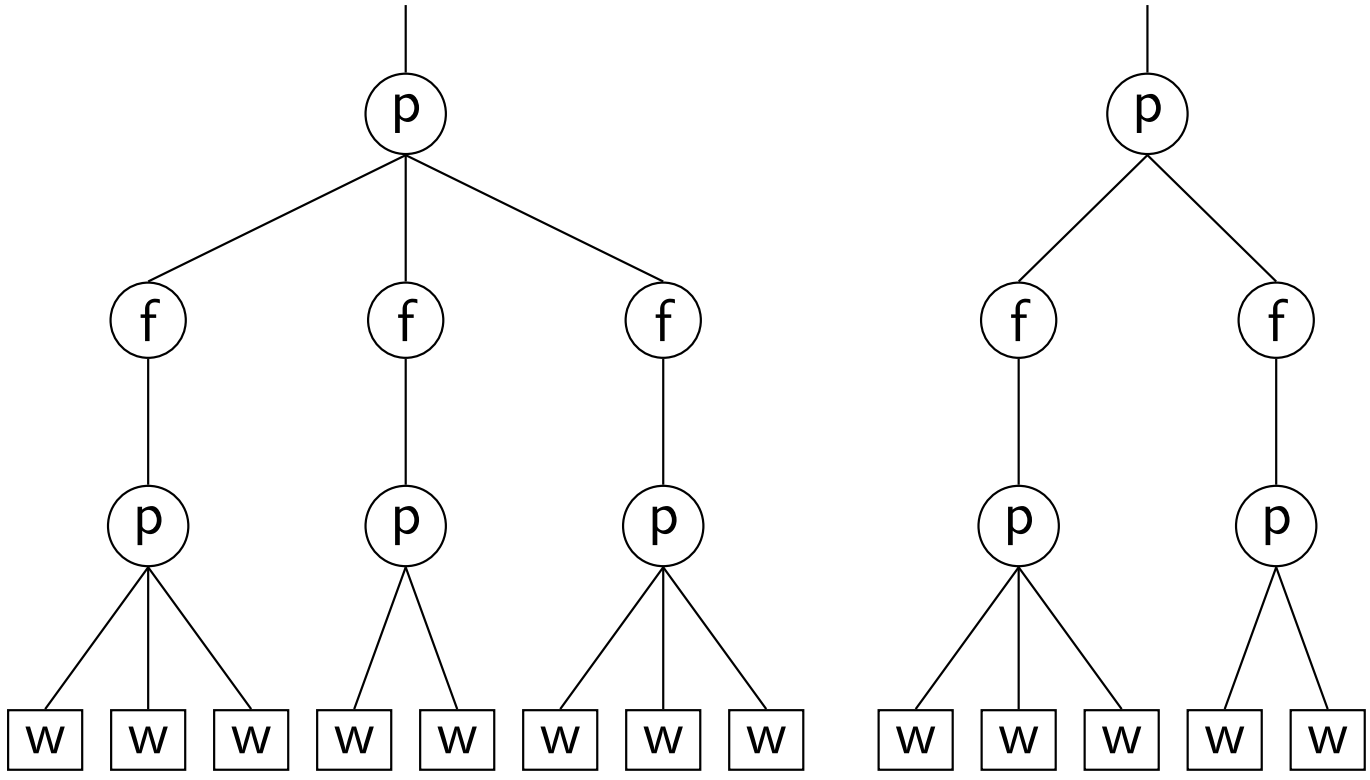
We obtain $f : [\![1, k+m]\!] \to [\![1, k+m+m']\!]$ and we can recode $j \in [\![1, k+m]\!]$ as $f(j)h(j)$ where $h(j)$ is the rank of $j \in f^{-1}\left[f(j)\right]$.
The sample representation R becomes of type $\{\{fh\}\wedge\}$. It can be split into $\{\{f\}\wedge\}\{\{h\}\wedge\}$ where the values of h corresponding to each value of f have been gathered on the right. Each $\{f\}$ describes the syntax of a sample image at the first level.
The whole compression process can be repeated on the first level sample syntax $\{\{f\}\wedge\}$ to get several levels of syntax organized into a syntax tree.

Syntax tree

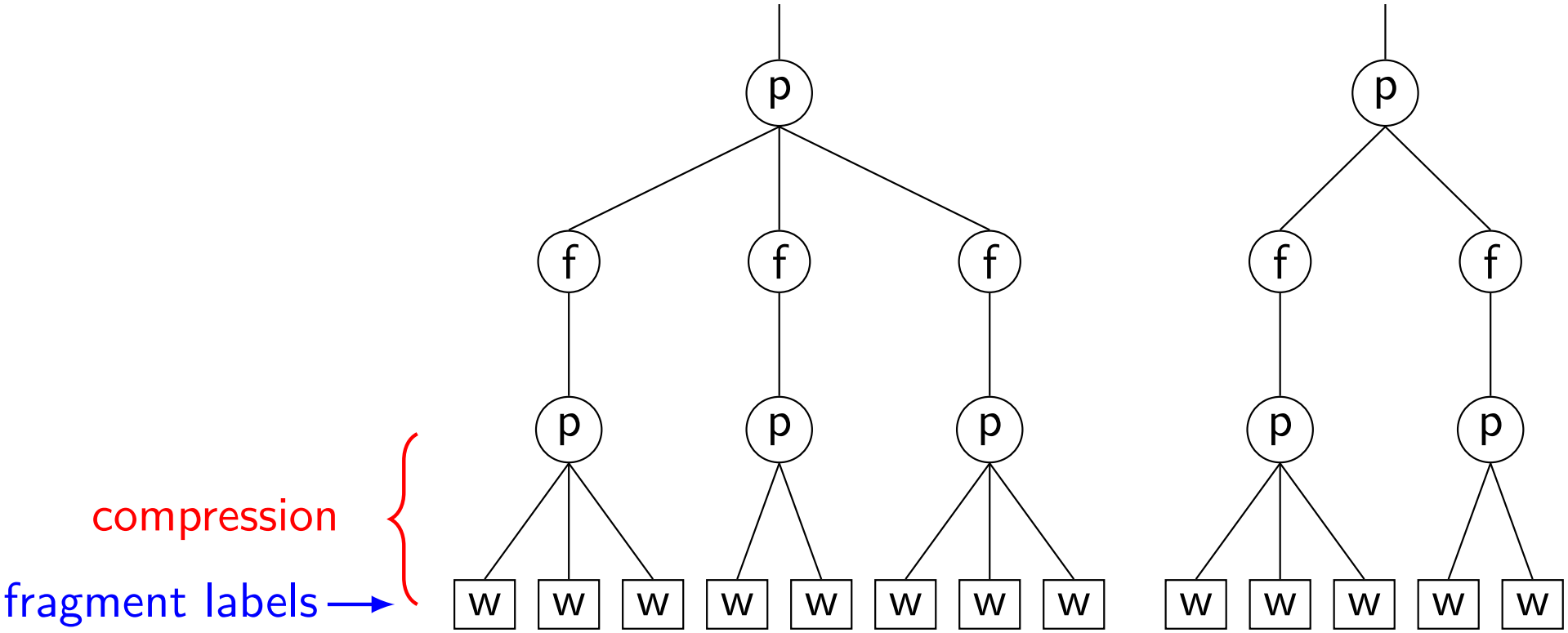


Syntax tree

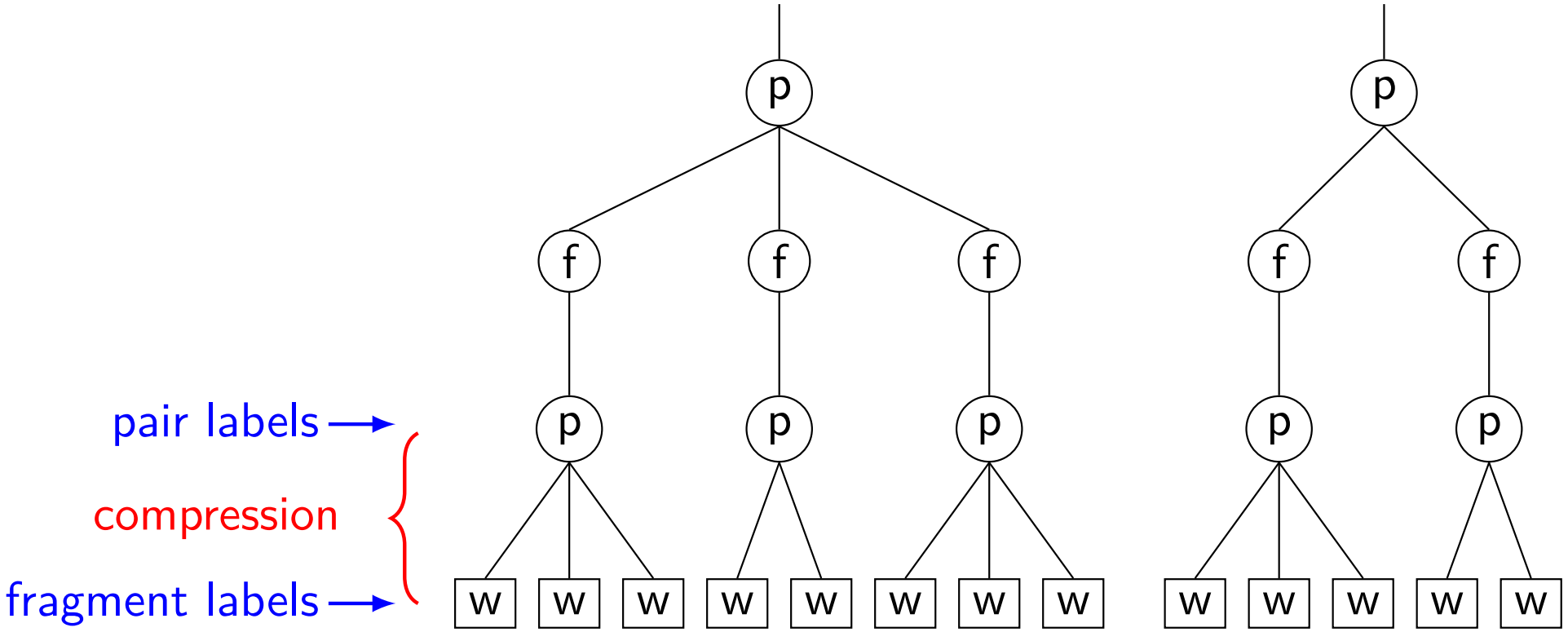
fragment labels ⟶



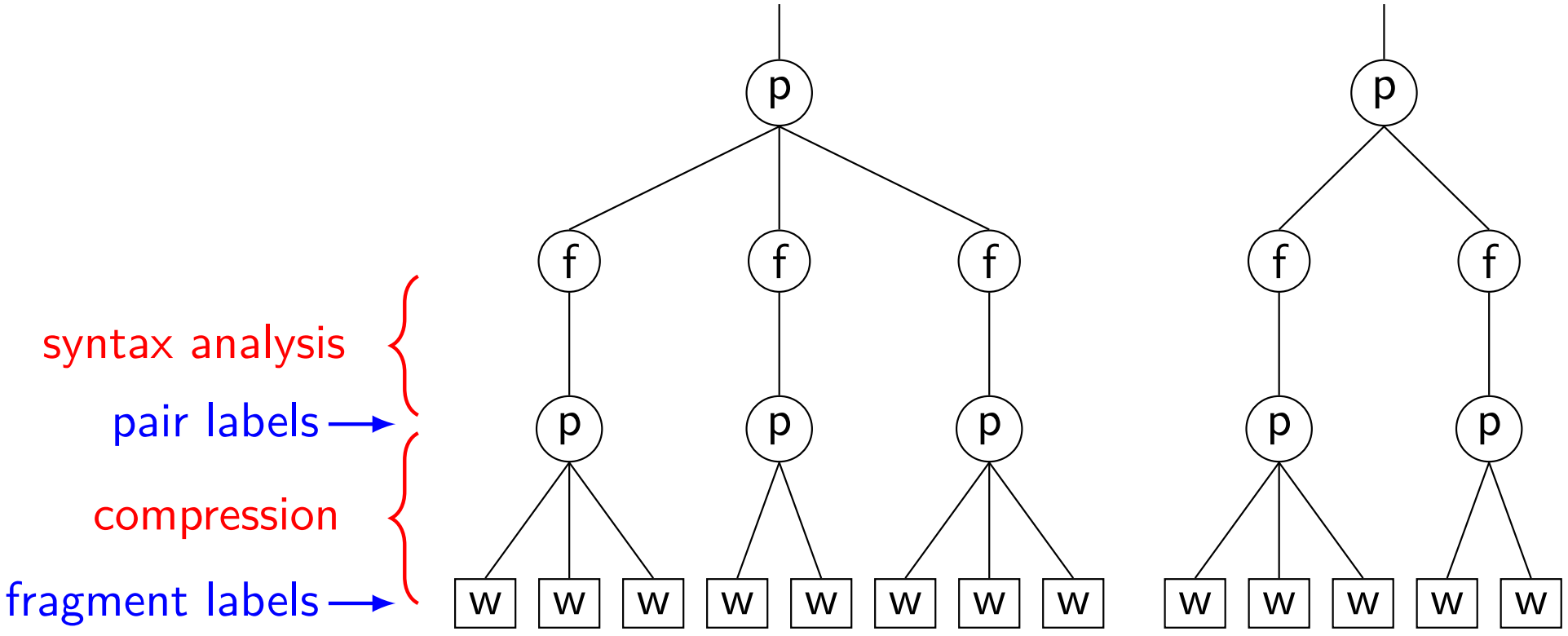
Syntax tree



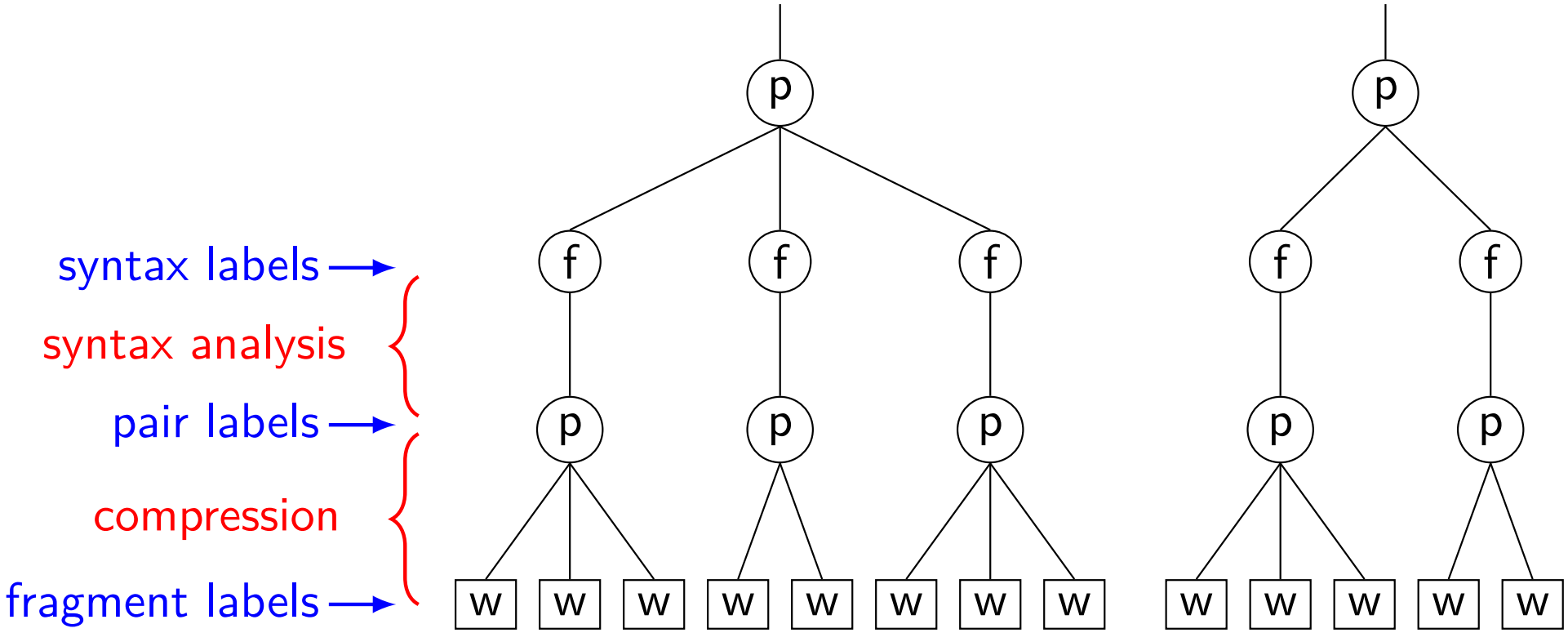
Syntax tree


pair labels
compression
fragment labels
p
f
f
f
p
p
p
w
w
w
w
w
w
w
w
p
f
f
p
p
w
w
w
w
w

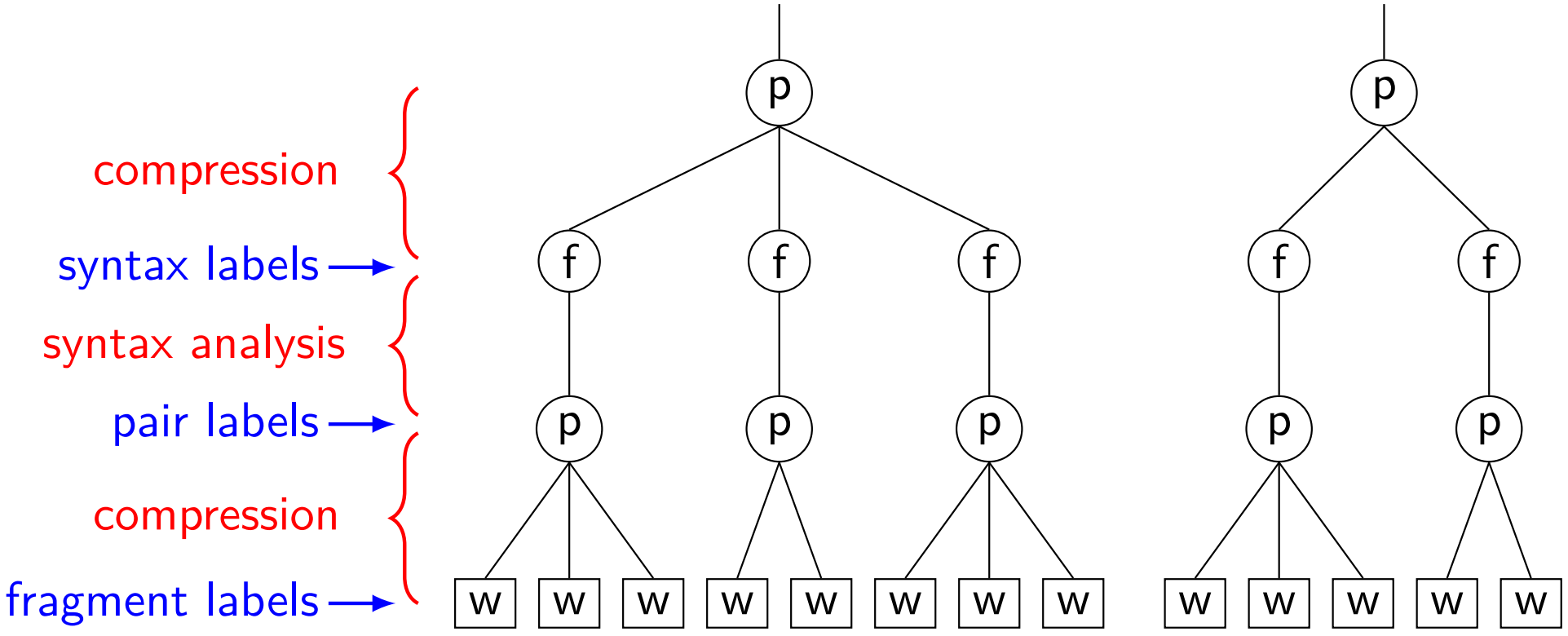
Syntax tree


syntax analysis
pair labels
compression
fragment labels
p
f
w

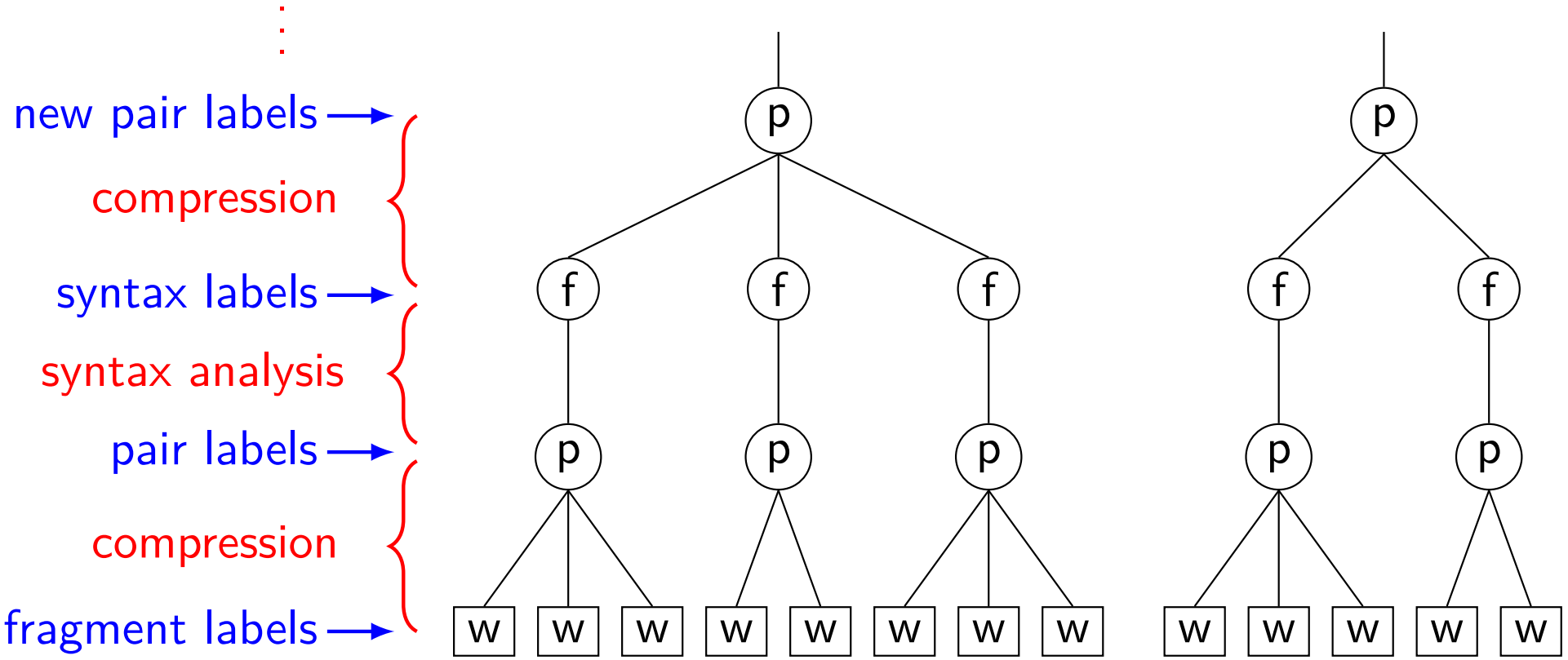
Syntax tree


syntax labels
syntax analysis
pair labels
compression
fragment labels
p
f
f
f
p
p
p
w
w
w
w
w
w
w
w
p
f
f
p
p
w
w
w
w
w

Syntax tree


compression
syntax labels
syntax analysis
pair labels
compression
fragment labels
p
f
f
f
p
p
p
w
w
w
w
w
w
w
w
p
f
f
p
p
w
w
w
w
w

Syntax tree


new pair labels
compression
syntax labels
syntax analysis
pair labels
compression
fragment labels
p
f
w

Lemma for linear k-means in a separable Hilbert space

- Let $(W_1, \dots, W_n)$ be n independent copies of $W \in \ell^2$.
- Assume that $\|W\|_\infty = \text{ess sup}\,\|W\| < +\infty$.
- Let $\Theta \subset \left(\ell^2\right)^k$ be bounded.
- Assume that $\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]} \langle \theta_j, W\rangle \in [a,b] \text{ for all } \theta \in \Theta\right) = 1$.
- For any $k \geq 2$, any $n \geq 2k$ and any $\delta \in]0,1[$, with probability at least $1-\delta$, for any $\theta \in \Theta$,

$$\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]} \langle \theta_j, W\rangle\right) \leq \overline{\mathbb{P}}_W\left(\min_{j\in[\![1,k]\!]} \langle \theta_j, W\rangle\right) + \left(\frac{\log(n/k)}{\log(2)}\sqrt{\frac{8\log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}}\right)\|\Theta\|\|W\|_\infty + \sqrt{\frac{(\sqrt{2}+1)\left(k(b-a)^2 + 2\log(ek)\|W\|_\infty^2\|\Theta\|^2\right)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}(b-a).$$

Lemma on the excess risk

- If $\theta^* \in \Theta$ is non random, with probability at least $1-\delta$,

$$\sup_{\theta\in\Theta}\left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right)\left(\min_{j\in[\![1,k]\!]}\langle\theta_j, W\rangle - \min_{j\in[\![1,k]\!]}\langle\theta_j^*, W\rangle\right)$$

$$\leq \left(\frac{\log(n/k)}{\log(2)}\sqrt{\frac{8\log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}}\right)\|\Theta\|\|W\|_\infty +$$

$$\sqrt{\frac{(\sqrt{2}+1)\left(k(b-a)^2 + 2\log(ek)\|W\|_\infty^2\|\Theta\|^2\right)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}(b-a).$$

- If $\widehat{\theta} \in \arg\min_{\theta\in\Theta}\overline{\mathbb{P}}\left(\min_{j\in[\![1,k]\!]}\langle\theta_j, W\rangle\right)$, we have the same bound for

$$\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]}\langle\widehat{\theta}_j, W\rangle\right) - \inf_{\theta\in\Theta}\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]}\langle\theta_j, W\rangle\right).$$

Lemma

In expectation with respect to the sample distribution

$$\mathbb{P}_{W_1,\dots,W_n}\Big[\mathbb{P}_W\Big(\min_{j\in[\![1,k]\!]}\langle\widehat{\theta}_j,W\rangle\Big)-\inf_{\theta\in\Theta}\mathbb{P}_W\Big(\min_{j\in[\![1,k]\!]}\langle\theta_j,W\rangle\Big)\Big]$$
$$\leq\left(\frac{\log(n/k)}{\log(2)}\sqrt{\frac{8\log(k)}{n}}+2\sqrt{\frac{\log(k)}{n}}\right)\|\Theta\|\|W\|_\infty$$
$$+\sqrt{\frac{(\sqrt{2}+1)\big(k(b-a)^2+2\log(ek)\|W\|_\infty^2\|\Theta\|^2\big)}{n}}.$$

Proposition

- Let $(X_1, \dots, X_n)$ be n independent copies of $X \in H$, some separable Hilbert space.
- Assume that $\mathbb{P}\big(\|X\| \leq B\big) = 1$, $n \geq 2k$ and $k \geq 2$.
- Consider an estimator $\widehat{C} \in \arg\min_{C \in H^k} \overline{\mathbb{P}}_X\big(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\big)$.

$$\mathbb{P}_{X_1,\dots,X_n}\Big[\mathbb{P}_X\big(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\big)\Big] \leq \inf_{C \in H^k} \mathbb{P}_X\big(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\big) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right)\sqrt{\frac{k\log(k)}{n}}}.$$

Improves on [Biau Devroye and Lugosi, 2008] $\mathcal{O}\big(k/\sqrt{n}\big)$.

Proof: Remark that $\|C_j\|^2 - 2\langle C_j, X\rangle = \langle \theta_j, W\rangle$, where $\theta_j = \big(C_j, \gamma^{-1}\|C_j\|^2 B^{-1}\big) \in H \times \mathbb{R}$ and $W = \big(-2X, \gamma B\big)$ and apply the previous lemma. Take $\gamma_{\text{opt}} = \sqrt{2}$.

Lemma. PAC-Bayesian inequality

- Consider $h : \mathcal{T} \times \mathcal{W} \to \mathbb{R}$.
- Consider a prior $\pi \in \mathcal{M}_+^1(\mathcal{T})$ on the parameter space $\mathcal{T}$.
- Let $\lambda > 0$ be a positive exponent.
- Let $(W_1, \ldots, W_n)$ be n independent copies of $W \in \mathcal{W}$.

$$\mathbb{P}_{W_1,\ldots,W_n}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_+^1(\mathcal{T})}\sup_{\eta\in\mathbb{N}}\left\{\int \min\left\{\eta, -\lambda\sum_{i=1}^n h(\theta, W_i)\right.\right.\right.\right.$$
$$\left.\left.\left.\left. - n\log\left[\mathbb{P}_W \exp\left[-\lambda h(\theta, W)\right]\right]\right\}\mathrm{d}\rho(\theta) - \mathcal{K}(\rho,\pi)\right\}\right]\right\} \leq 1.$$

- Let $\rho_{\theta'|\theta} = \bigotimes_{j=1}^{k} \left(\bigotimes_{i\in\mathbb{N}} \mathcal{N}(\theta_{j,i}, \beta^{-1}) \right) : (\mathbb{R}^{\mathbb{N}})^k \to \mathcal{M}_+^1\big((\mathbb{R}^{\mathbb{N}})^k\big)$.
- Let

$$\langle \theta, w \rangle = \begin{cases} \lim\limits_{s\to+\infty} \sum\limits_{i=0}^{s} \theta_i w_i, & \text{when } \overline{\lim\limits_{s\to+\infty}} \sum\limits_{i=0}^{s} \theta_i w_i = \underline{\lim\limits_{s\to+\infty}} \sum\limits_{i=0}^{s} \theta_i w_i \in \mathbb{R}, \\ 0, & \text{otherwise} \end{cases}$$

  be a non bilinear but measurable extension of the scalar product from ℓ^2 to $\mathbb{R}^{\mathbb{N}}$.
- Introduce $f(\theta, w) = \min\limits_{j\in[\![1,k]\!]} \langle \theta_j, w \rangle, \quad \theta \in (\mathbb{R}^{\mathbb{N}})^k, w \in \mathbb{R}^{\mathbb{N}}$
- and the centered loss function $\overline{f}(\theta, w) = f(\theta, w) - \mathbb{P}_W\big(f(\theta, W)\big)$.

- Write

$$(\mathbb{P}_W - \overline{\mathbb{P}}_W) f(\theta, W) = (\mathbb{P}_W - \overline{\mathbb{P}}_W)(\delta_{\theta'|\theta} - \underbrace{\rho_{\theta'|\theta}}_{\text{small perturbation}}) f(\theta', W)$$

$$+ \sum_{p=1}^{H} (\mathbb{P}_W - \overline{\mathbb{P}}_W)(\rho_{\theta'|\theta}^{2^{p-1}} - \underbrace{\rho_{\theta'|\theta}^{2^p}}_{\text{chain of intermediate scales}}) f(\theta', W)$$

$$+ (\mathbb{P}_W - \overline{\mathbb{P}}_W) \underbrace{\rho_{\theta'|\theta}^{2^H}}_{\text{big perturbation}} f(\theta', W).$$

- Remark that

$$(\delta_{\theta'|\theta} - \rho_{\theta'|\theta}) f(\theta', W) = \rho_{\theta'|\theta} \left(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle - \min_{j \in [\![1,k]\!]} \langle \theta'_j, W \rangle \right)$$

$$\leq \rho_{\theta'|\theta} \left(\max_{j \in [\![1,k]\!]} \underbrace{\langle \theta_j - \theta'_j, W \rangle}_{\text{Gaussian}/\rho} \right) \leq \sqrt{2 \log(k)/\beta}\, \|W\|_\infty.$$

Chaining inequalities

- From the PAC-Bayesian inequality applied to $h(\theta, w) = \left(\delta_{\theta' \mid \theta} - \rho_{\theta' \mid \theta}\right) \overline{f}(\theta', w)$,

$$\mathbb{P}_{W_1,\dots,W_n} \Bigg\{ \exp \sup_{\theta \in (\ell^2)^k} \Bigg[n\lambda \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \left(\rho_{\theta' \mid \theta} - \rho^2_{\theta' \mid \theta}\right) f(\theta', W) - n\rho_{\theta' \mid \theta} \left[\log \left(\mathbb{P}_W \left[\exp \left(-\lambda \left(\delta_{\theta'' \mid \theta'} - \rho_{\theta'' \mid \theta'}\right) \overline{f}(\theta'', W) \right) \right] \right) \right] - \frac{\beta \|\theta\|^2}{2} \Bigg] \Bigg\} \leq 1.$$

- This gives

$$\mathbb{P}_{W_1,\dots,W_n} \left[\sup_{\theta \in \Theta} \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \left(\rho_{\theta' \mid \theta} - \rho^2_{\theta' \mid \theta}\right) f(\theta', W) \right] \leq \frac{4\lambda}{\beta} \log(k) \|W\|_\infty^2 + \frac{\beta \|\Theta\|^2}{2n\lambda} \underset{\lambda_{\text{opt}}}{=} \|W\|_\infty \|\Theta\| \sqrt{\frac{8 \log(k)}{n}}.$$

Bounding the biggest perturbation

- Consider $\psi(x) = \begin{cases} \log\left(1+x+x^2/2\right), & x \geq 0, \\ -\log(1-x+x^2/2), & x \leq 0, \end{cases}$
- and $\widetilde{f}(\theta, W) = f(\theta, W) - \dfrac{a+b}{2}$.
- Remark that

$$\left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right)\rho_{\theta'|\theta} f(\theta', W) =$$

$$\rho_{\theta'|\theta}\left[\mathbb{P}_W \widetilde{f}(\theta', W) - \overline{\mathbb{P}}_W\left(\lambda^{-1}\psi\left[\lambda \widetilde{f}(\theta', W)\right]\right)\right]$$

$$+ \underbrace{\rho_{\theta'|\theta}\overline{\mathbb{P}}_W\left[\lambda^{-1}\psi\left[\lambda \widetilde{f}(\theta', W)\right] - \widetilde{f}(\theta', W)\right]}_{\leq \frac{\lambda}{2(1+\sqrt{2})}\left[(b-a)^2/4 + 2\log(ek)\|W\|_\infty^2/\beta\right]}.$$

since $|x - \psi(x)| \leq \frac{x^2}{4(1+\sqrt{2})}, \quad x \in \mathbb{R}.$

PAC-Bayesian inequality with an influence function

- Take $h(\theta, w) = \lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta, w)\right]$ to obtain

$$\mathbb{P}_{W_1,\ldots,W_n}\left\{\sup_{\theta\in\Theta}\exp\left[-n\lambda\rho_{\theta'\,|\,\theta}\overline{\mathbb{P}}_W\left(\lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta', W)\right]\right)\right.\right.$$
$$\left.\left. - n\rho_{\theta'\,|\,\theta}\left[\log\left(\mathbb{P}_W\left[\exp\left(-\psi\left[\lambda\widetilde{f}(\theta', W)\right]\right)\right]\right)\right] - \frac{\beta\|\theta\|^2}{2}\right]\right\} \leq 1.$$

- Use $\psi(x) \leq \log\left(1 + x + x^2/2\right), \quad x \in \mathbb{R}$ to deduce

$$\mathbb{P}_{W_1,\ldots,W_n}\left\{\sup_{\theta\in\Theta}\rho_{\theta'\,|\,\theta}\left[\mathbb{P}_W\left(\widetilde{f}(\theta', W)\right) - \overline{\mathbb{P}}_W\left(\lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta', W)\right]\right)\right]\right\}$$
$$\leq \lambda\left[(b-a)^2/4 + 2\log(ek)\|W\|_\infty^2/\beta\right] + \frac{\beta\|\Theta\|^2}{2n\lambda}.$$

Putting all together

- For the biggest perturbation we get

$$\mathbb{P}_{W_1,\dots,W_n}\left\{\sup_{\theta\in\Theta}\left(\mathbb{P}_W-\overline{\mathbb{P}}_W\right)\rho_{\theta'|\theta}^{2^H}f(\theta',W)\right\}$$

$$\underset{\lambda_{\text{opt}}}{\leq}\sqrt{\frac{(\sqrt{2}+1)\left(2^{-H}\beta(b-a)^2+8\log(ek)\|W\|_\infty^2\right)\|\Theta\|^2}{4n}}.$$

- Putting all together

$$\mathbb{P}_{W_1,\dots,W_n}\left\{\sup_{\theta\in\Theta}\left(\mathbb{P}_W-\overline{\mathbb{P}}_W\right)f(\theta,W)\right\}\leq 2\sqrt{2\log(k)/\beta}\|W\|_\infty$$

$$+\sqrt{\frac{(\sqrt{2}+1)\left(2^{-H}\beta(b-a)^2+8\log(ek)\|W\|_\infty^2\right)\|\Theta\|^2}{4n}}$$

$$+H\|W\|_\infty\|\Theta\|\sqrt{\frac{8\log(k)}{n}}.$$

- Choose $\beta=2n\|\Theta\|^{-2}$ and $H=\left\lfloor\log(n/k)/\log(2)\right\rfloor$.

According to the bounded difference inequality

With probability at least $1-\delta$

$$\sup_{\theta\in\Theta}\left(\mathbb{P}_W-\overline{\mathbb{P}}_W\right)f(\theta,W)$$

$$\leq \mathbb{P}_{W_1,\dots,W_n}\left\{\sup_{\theta\in\Theta}\left(\mathbb{P}_W-\overline{\mathbb{P}}_W\right)f(\theta,W)\right\}+\sqrt{\frac{2\log(\delta^{-1})}{n}}(b-a).$$

Generalization bounds for fragmentation

- Consider the fragmentation model

$$\mathcal{M}(\mathcal{S}) = \Big\{ (B,C)_{j=1}^{k} : B_j \subset [\![1,k]\!], C_j \in [-a,a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \Big\},$$

  where $\mathcal{T}_{B,K} = \big\{ A \subset \mathcal{T}_B : |A| \leq K \big\}$.

- Remark that $\log\big(|\mathcal{T}_{B,K}|\big) \leq K \log\left(\frac{ek}{K}\right)$.

- Consider the risk function

$$\mathcal{R}(B,C) = \underbrace{\mathcal{D}(B,C)}_{\stackrel{\text{def}}{=} d^{-1}\mathbb{P}_X\big(\min_{A\in\mathcal{T}_{B,K}} \|X - C_A\|^2\big)} - d^{-1}\mathbb{P}_X\big(\|X\|^2\big)$$

$$= d^{-1}\mathbb{P}_X\big[\min_{A\in\mathcal{T}_{B,K}} \big(\|C_A\|^2 - 2\langle X, C_A\rangle\big)\big].$$

- Assume that $\mathbb{P}\big(\max_{s\in[\![1,d]\!]} |X_s| \leq a\big) = 1$.

Proposition

Assume that $k \geq 2$, $K \geq 1$, $\mathcal{S} \in [1, k]$, $n \geq 2\mathcal{S}K$ and $\delta \in]0, 1[$. With probability at least $1 - \delta$, for any $(B, C) \in \mathcal{M}(\mathcal{S})$,

$$\begin{aligned}\mathcal{R}(B, C) - \overline{\mathcal{R}}(B, C) \leq a^2 \Bigg(& \frac{\sqrt{10}\log(n\mathcal{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} \\ & + 2\sqrt{10}\sqrt{\frac{\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} + \sqrt{\frac{4(\sqrt{2}+1)\big(9+5\log(|\mathcal{T}_{B,K}|)\big)\mathcal{S}K}{n}} \\ & + 2\sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{n}} \Bigg) \\ \leq a^2 \mathcal{O}\Bigg(& \log\left(\frac{n}{\mathcal{S}K}\right) \sqrt{\frac{\mathcal{S}K^2\log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \Bigg).\end{aligned}$$

This is based on a lemma for linear fragmentation similar to the linear k-means lemma.

Conclusion

- For a given empirical distortion level $\mathcal{D}(\overline{X}, B, C)$, we should minimize $\mathcal{S} = d^{-1}|B|$, to get the best possible generalization bound, similarly to what was suggested by the compression approach.
- We can take advantage of this generalization bound to compute the fragmentation code book (B, C) on a subsample of the data base to be analyzed.

Proposition

Let $(\widehat{B},\widehat{C}) \in \arg\min_{(B,C)\in\mathcal{M}(\mathcal{S})} \overline{\mathcal{D}}(B,C) + \left(\frac{\log(n\mathcal{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} + 2\sqrt{\frac{\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} \right) \sqrt{10}a^2 + \sqrt{\frac{4(\sqrt{2}+1)\big(4+5\log(e|\mathcal{T}_{B,K}|)\big)\mathcal{S}K}{n}}\, a^2.$

With probability at least $1-\delta$,

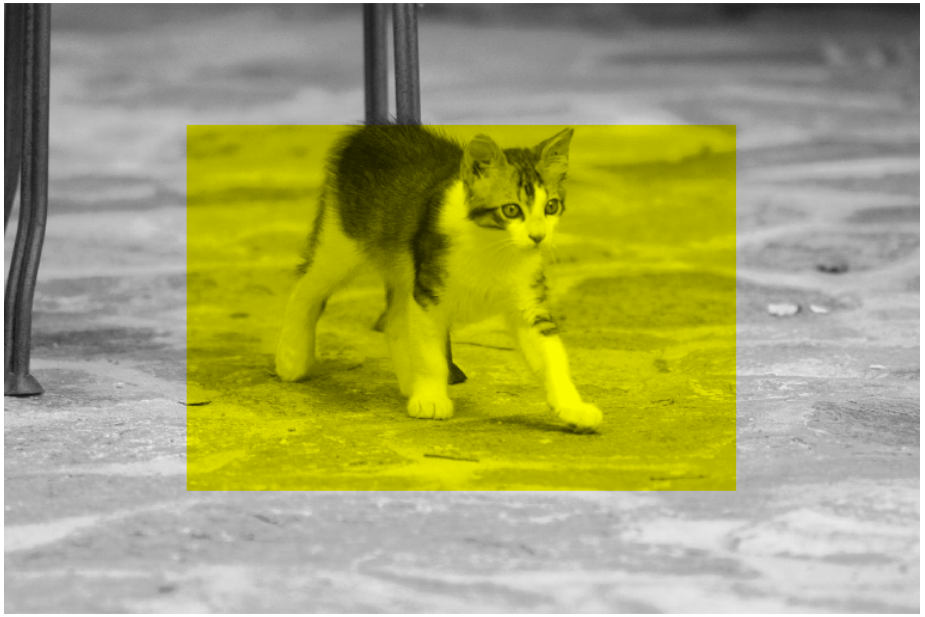
$$\mathcal{D}(\widehat{B},\widehat{C}) \leq \inf_{(B,C)\in\mathcal{M}(\mathcal{S})} \mathcal{D}(B,C) + \left(\frac{\log(n\mathcal{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} + 2\sqrt{\frac{\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}} \right) \sqrt{10}a^2 + \sqrt{\frac{4(\sqrt{2}+1)\big(4+5\log(e|\mathcal{T}_{B,K}|)\big)\mathcal{S}K}{n}} + 4\sqrt{\frac{k(k-1)\log(2)+2\log(\delta^{-1})}{n}}\, a^2.$$

Application to images: a small example

Some images to analyze

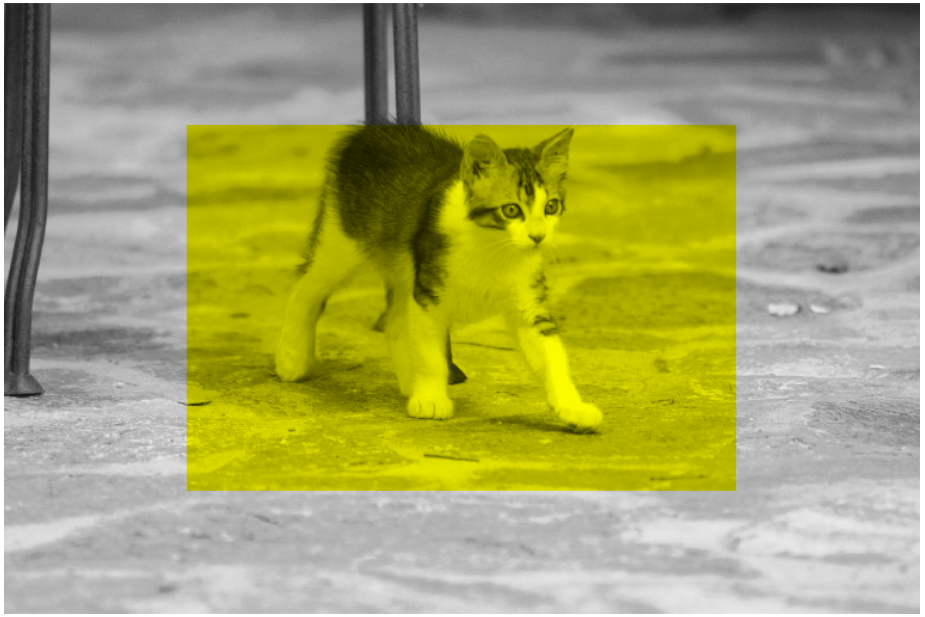
Application to images: a small example

Some images to analyze



Application to images: a small example

Some images to analyze



Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample


300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

$\ldots$

300 extracted patches of size 300×300 from the three previous yellow frames.

Application to images: extraction of the training sample

300 extracted patches of size 300×300 from the three previous yellow frames.

Fragmentation parameters

- Threshold for the distortion: $\alpha = 0.1 \times \frac{1}{d} \sum_{s=1}^{d} \mathbf{Var}\left(\overline{\mathbb{P}}_{X_s}\right)$
- Number of fragments: 2000
- Fragmentation criterion: looking for indexes that maximize

$$2 \underbrace{|B_j|}_{\substack{\text{size of} \\ \text{fragment } j}} - \underbrace{\sum_{i=1}^{n} \mathbb{1}(j \in A_i)}_{\substack{\text{number of images} \\ \text{sharing fragment } j}} .$$

Fragmentation: image approximation

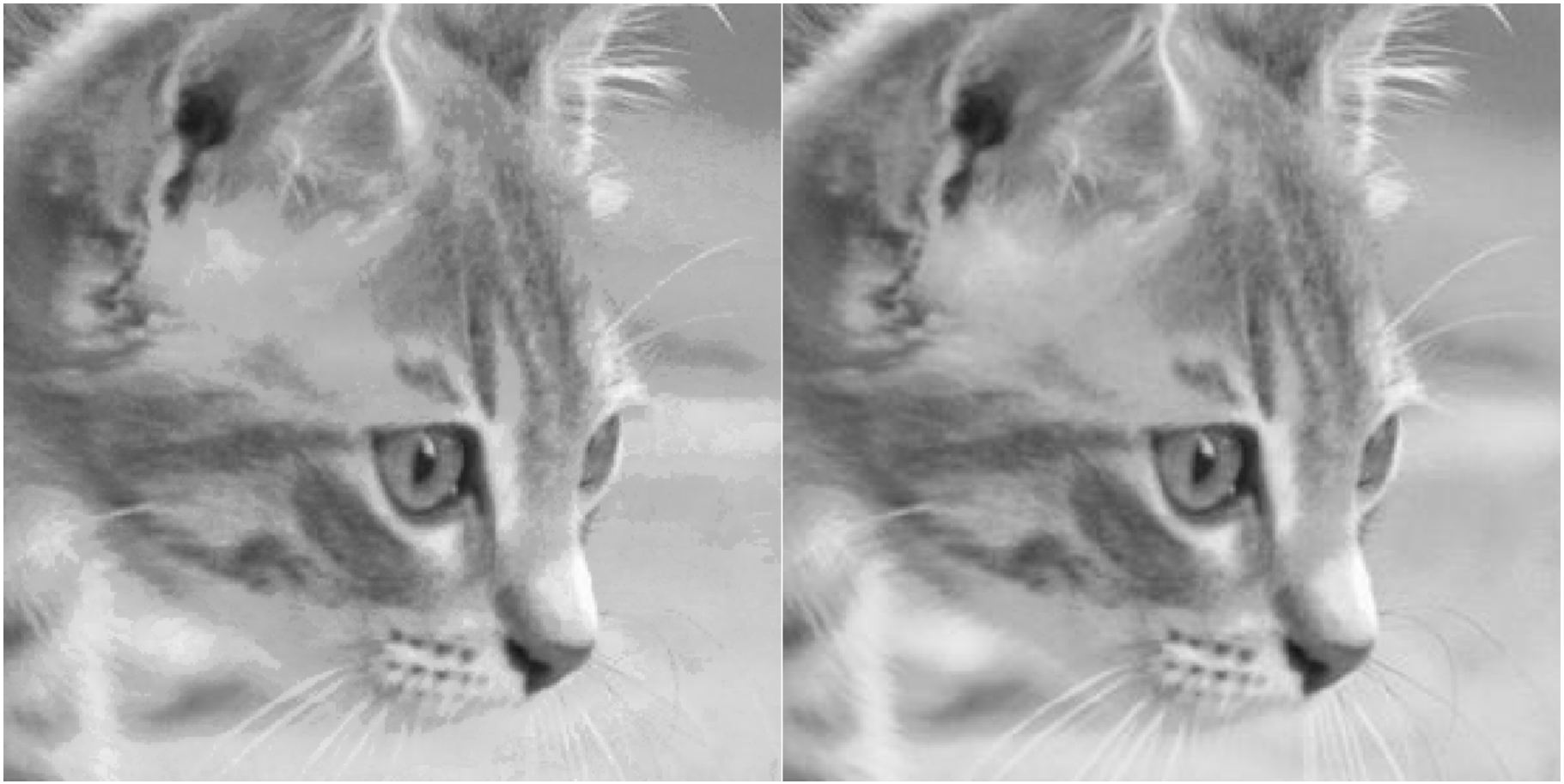


Figure: Image approximation resulting from the fragmentation vs original image.

Fragmentation: image approximation

Figure: Image approximation resulting from the fragmentation vs original image.

Fragmentation: image approximation

Figure: Image approximation resulting from the fragmentation vs original image.

Fragmentation: visualization of the fragments

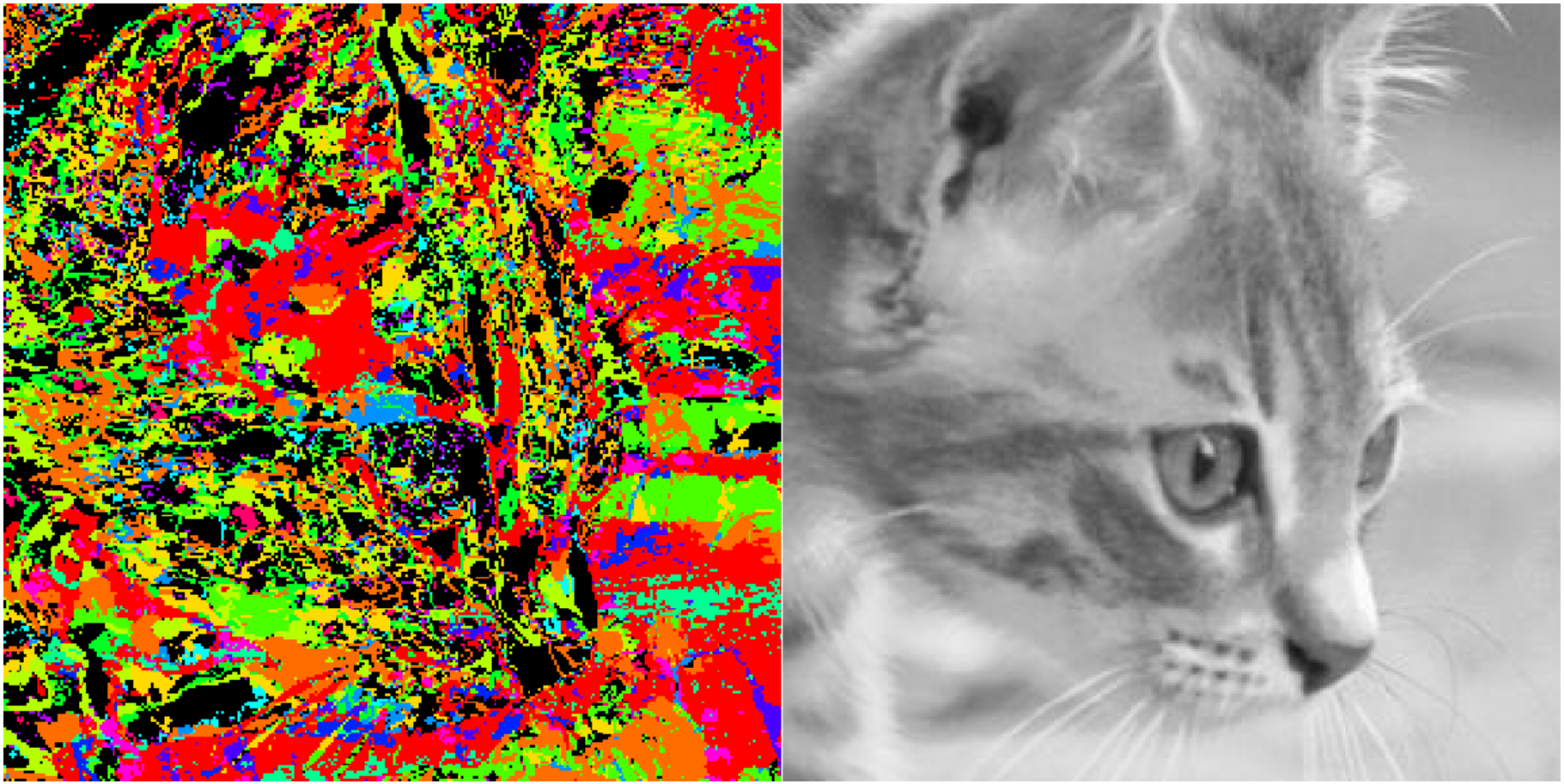


Figure: Visualization of the fragments.

Fragmentation: visualization of the fragments

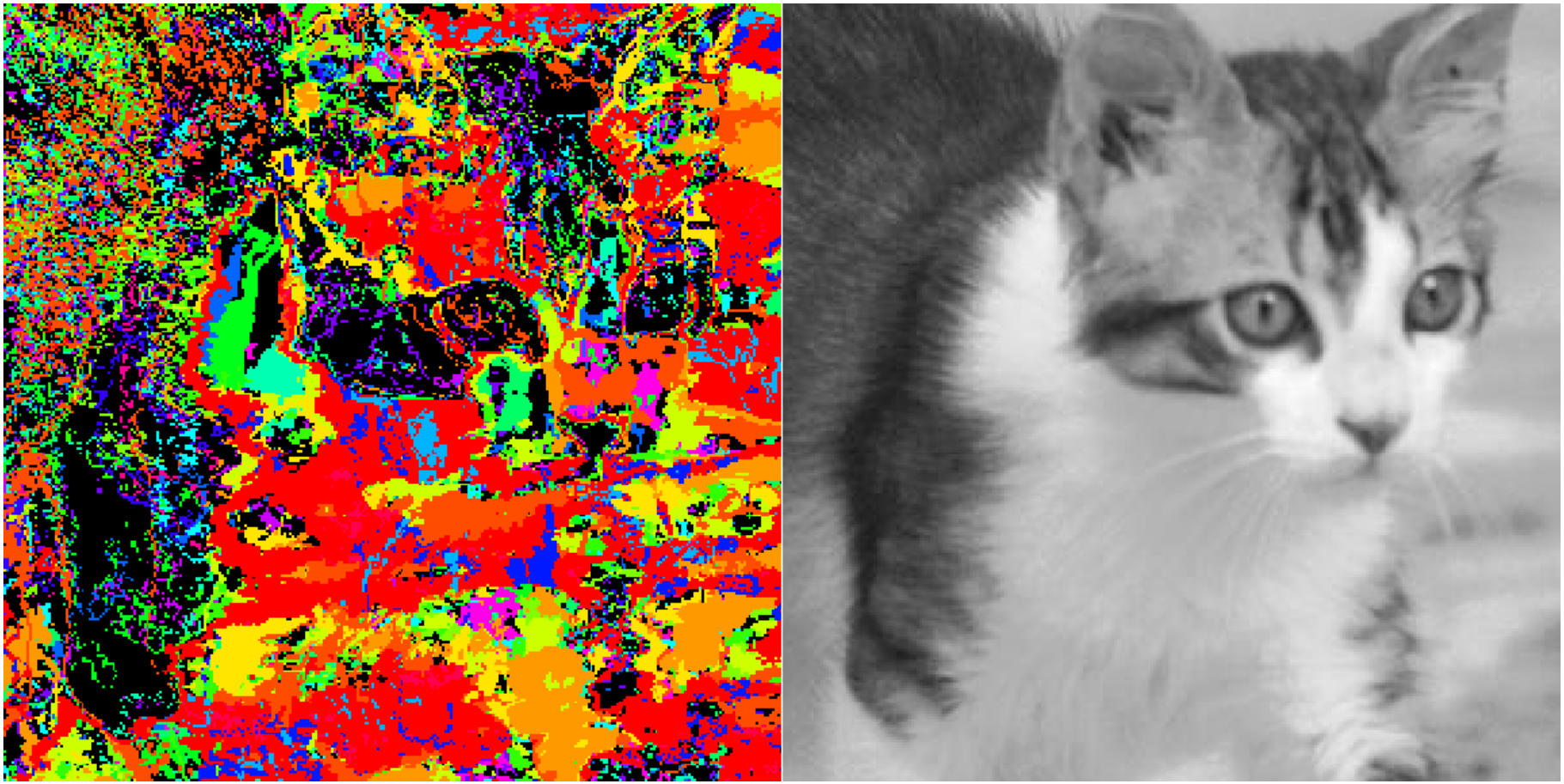


Figure: Visualization of the fragments.

Fragmentation: visualization of the fragments

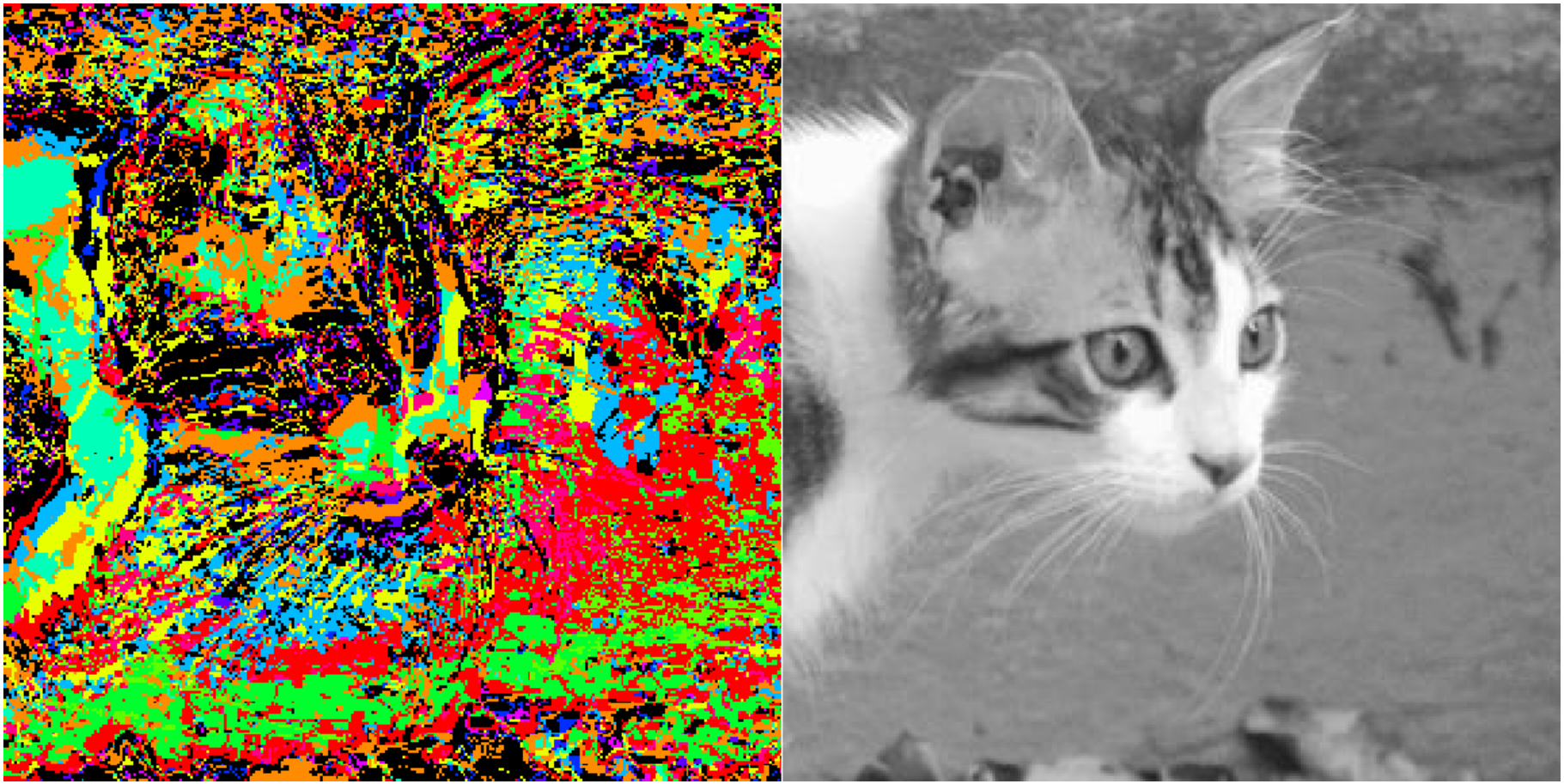


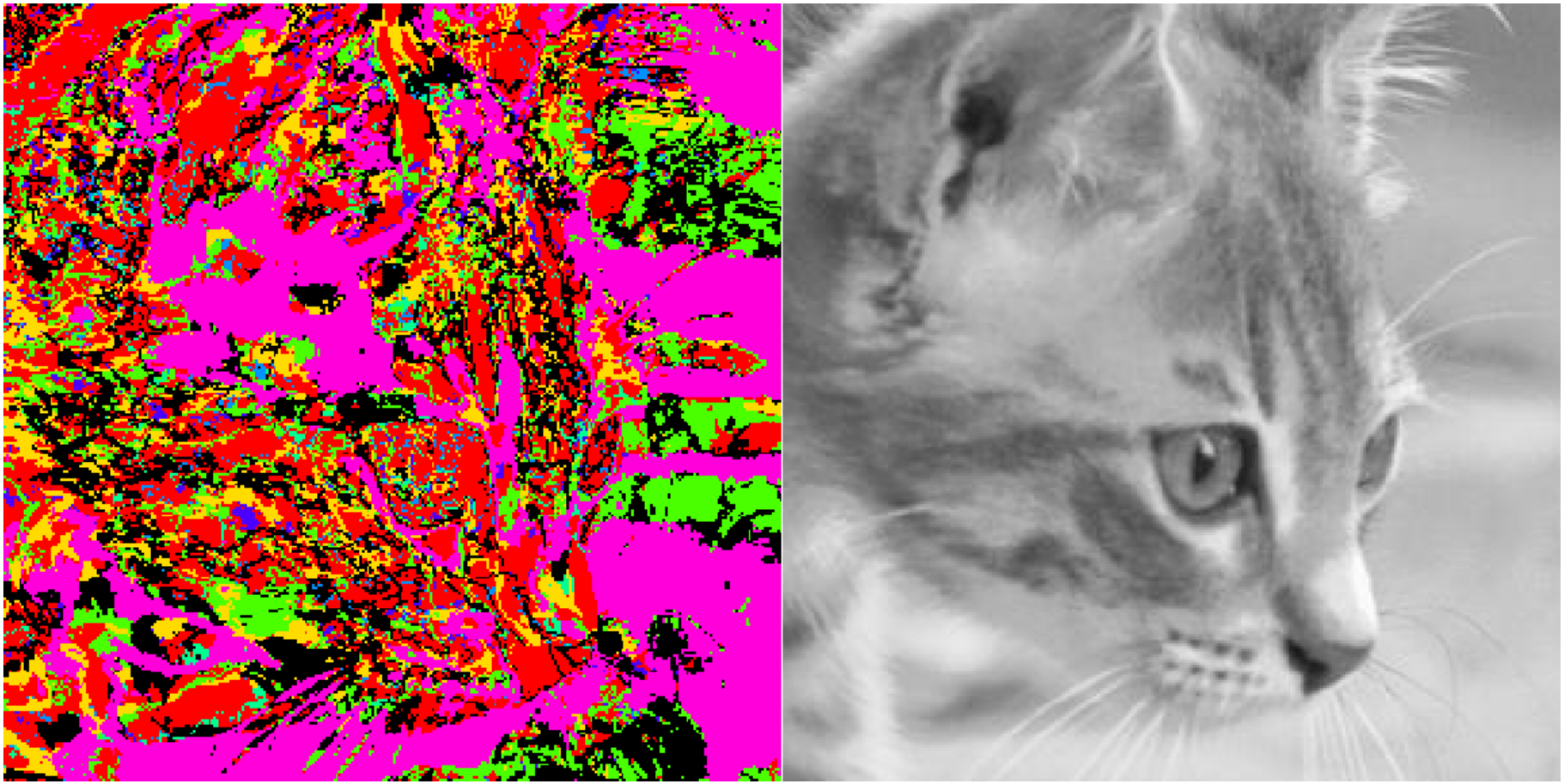
Figure: Visualization of the fragments.

Merging and Syntax parameters

- Maximum number of merged/compressed fragments: 300
- Maximum number of syntax labels: 300
- Iterate the merging step and syntax step until no merging is performed: 14 syntax steps.

Syntax labels

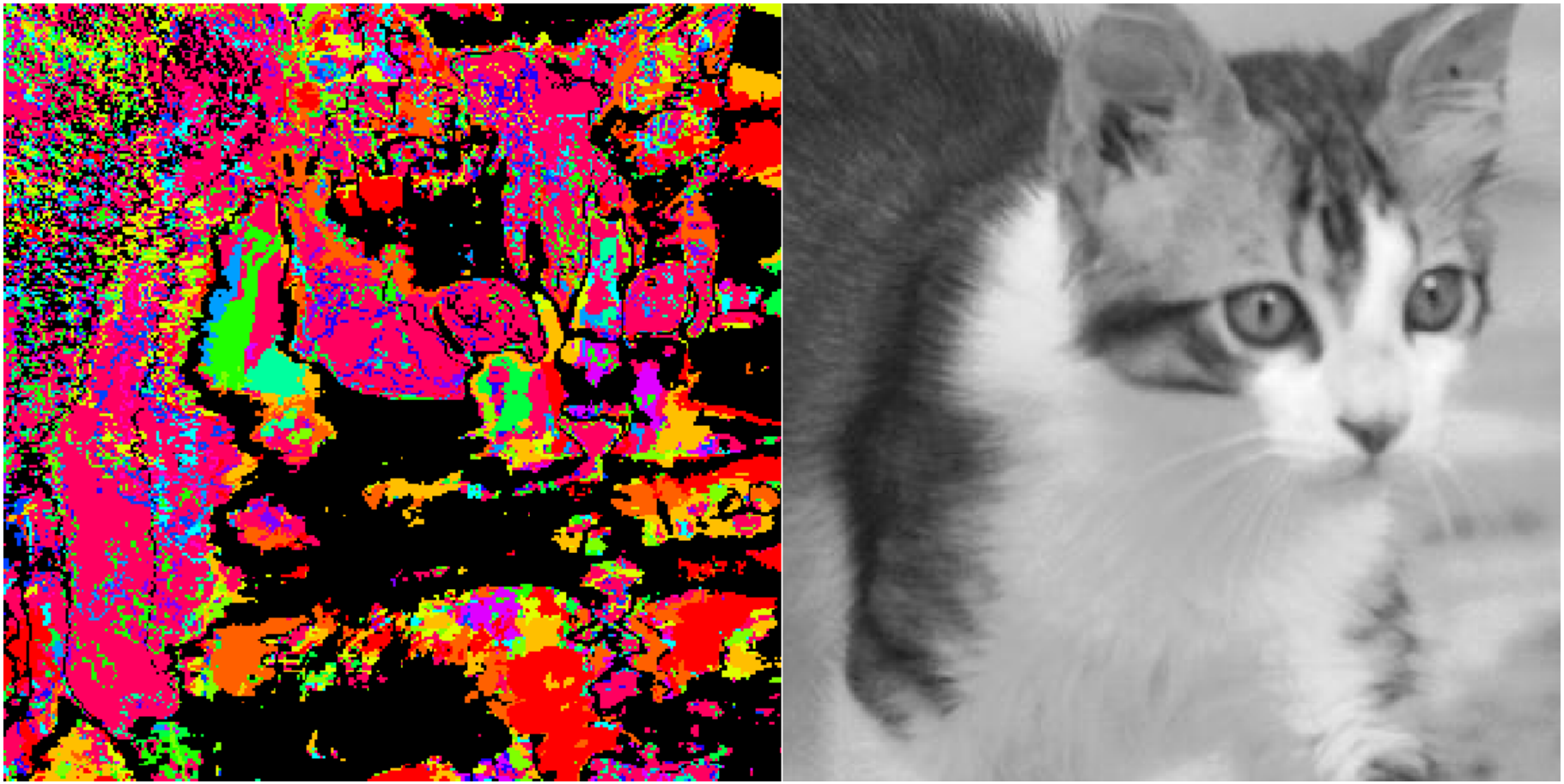
Syntax labels at highest level.



Compared to the result of fragmentation, some simplifications can be observed.

Syntax labels

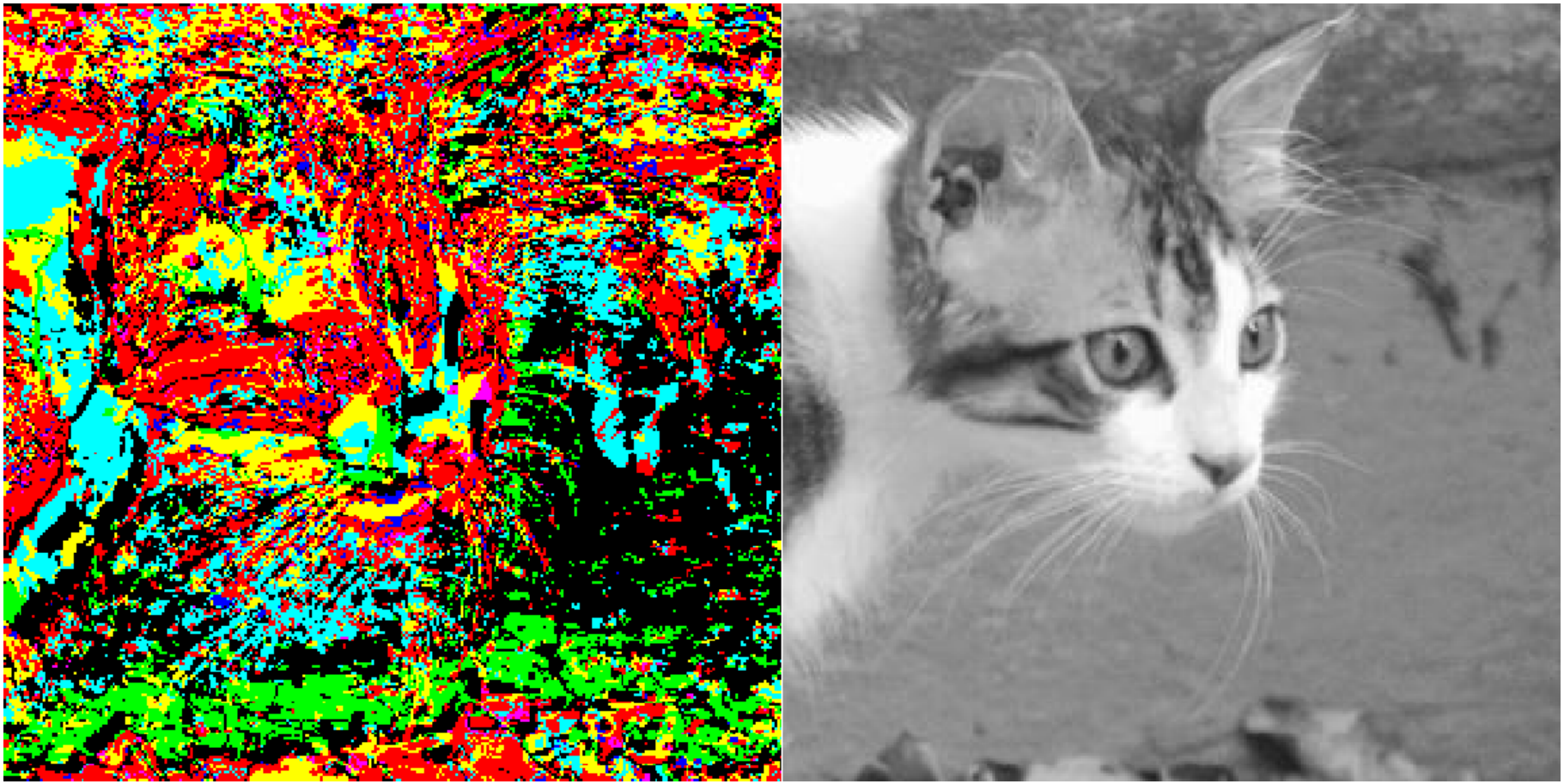
Syntax labels at highest level.



Compared to the result of fragmentation, some simplifications can be observed.

Syntax labels

Syntax labels at highest level.



Compared to the result of fragmentation, some simplifications can be observed.

Identifying translations

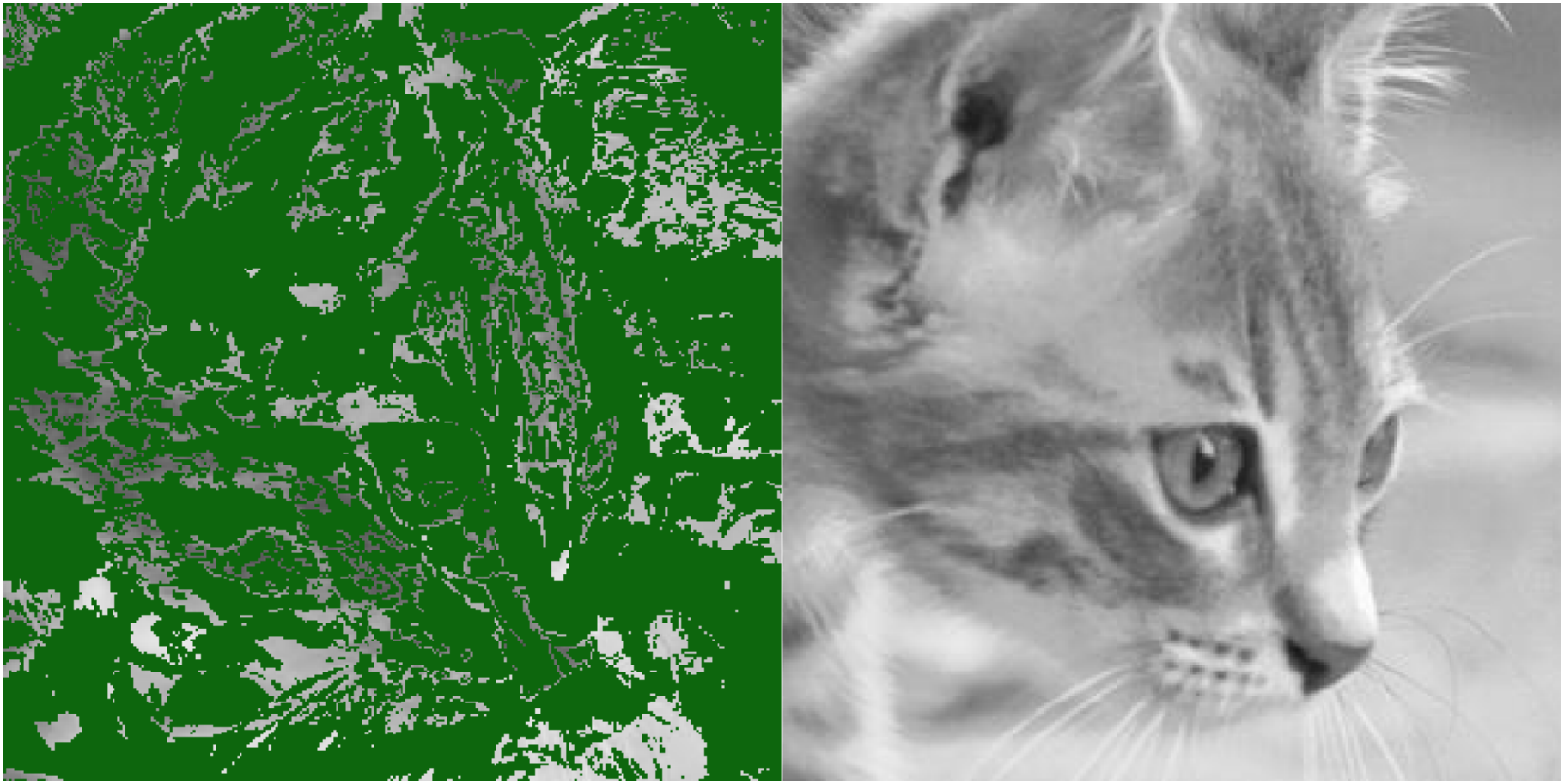


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

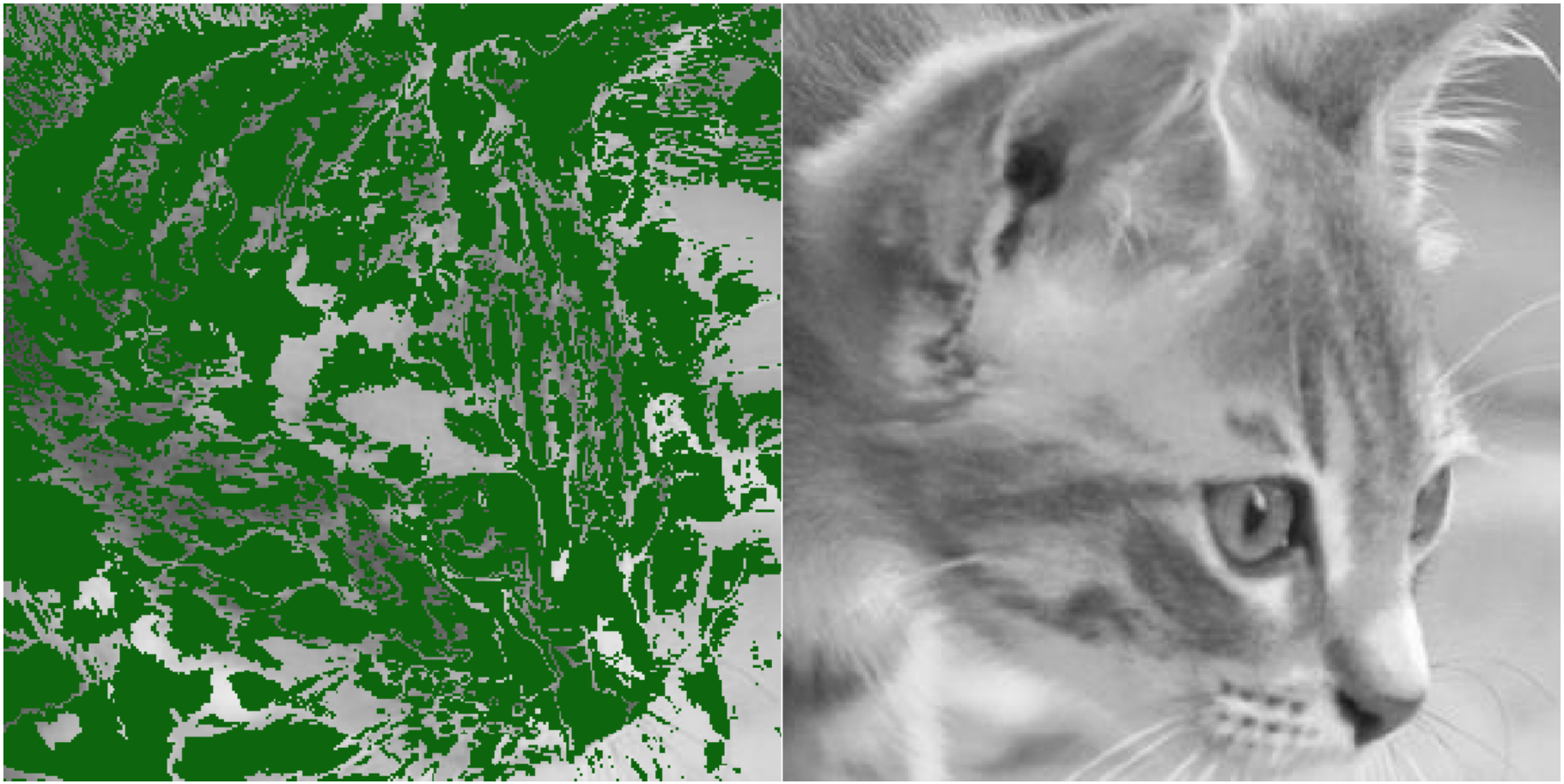


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

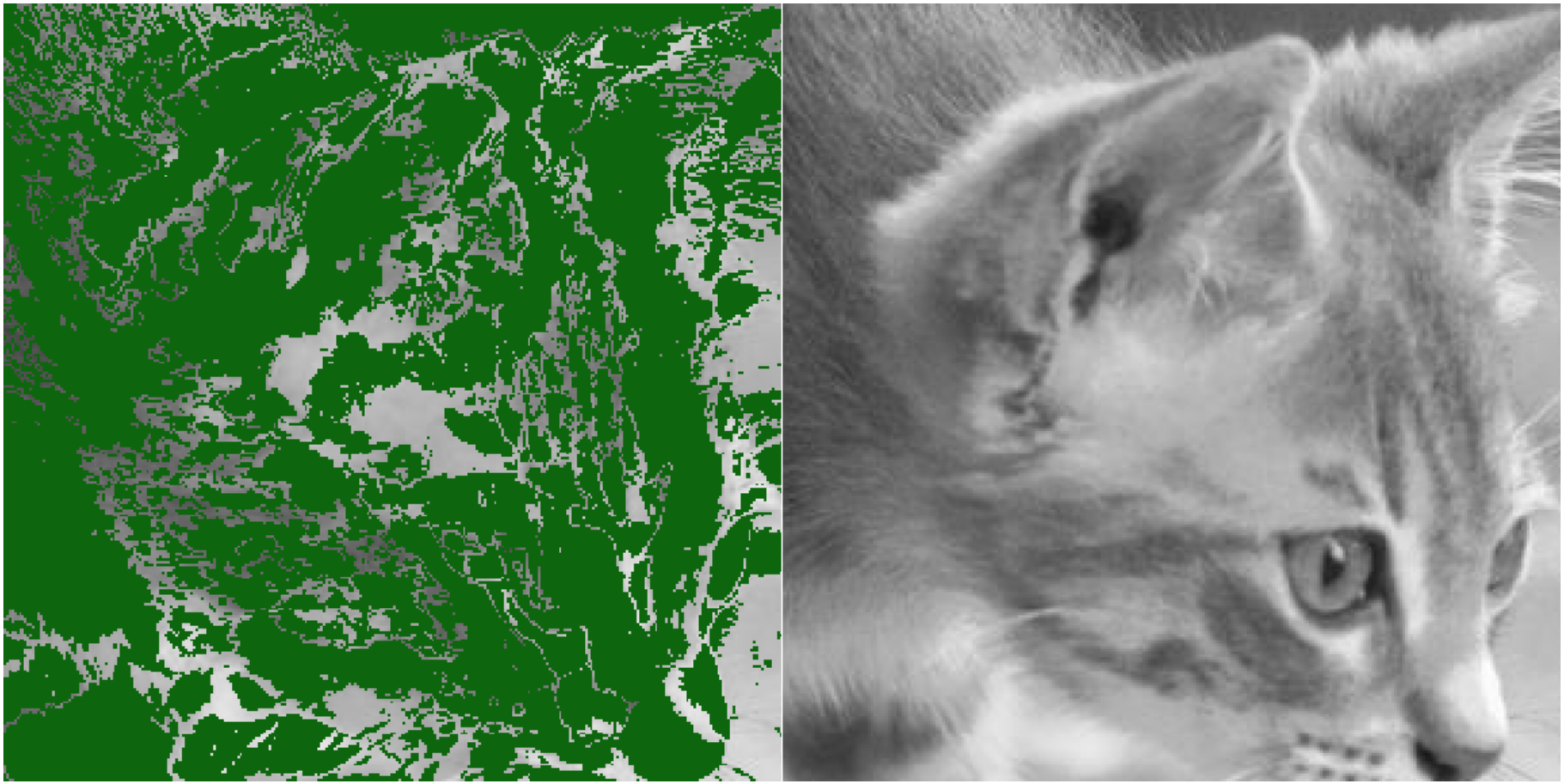


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

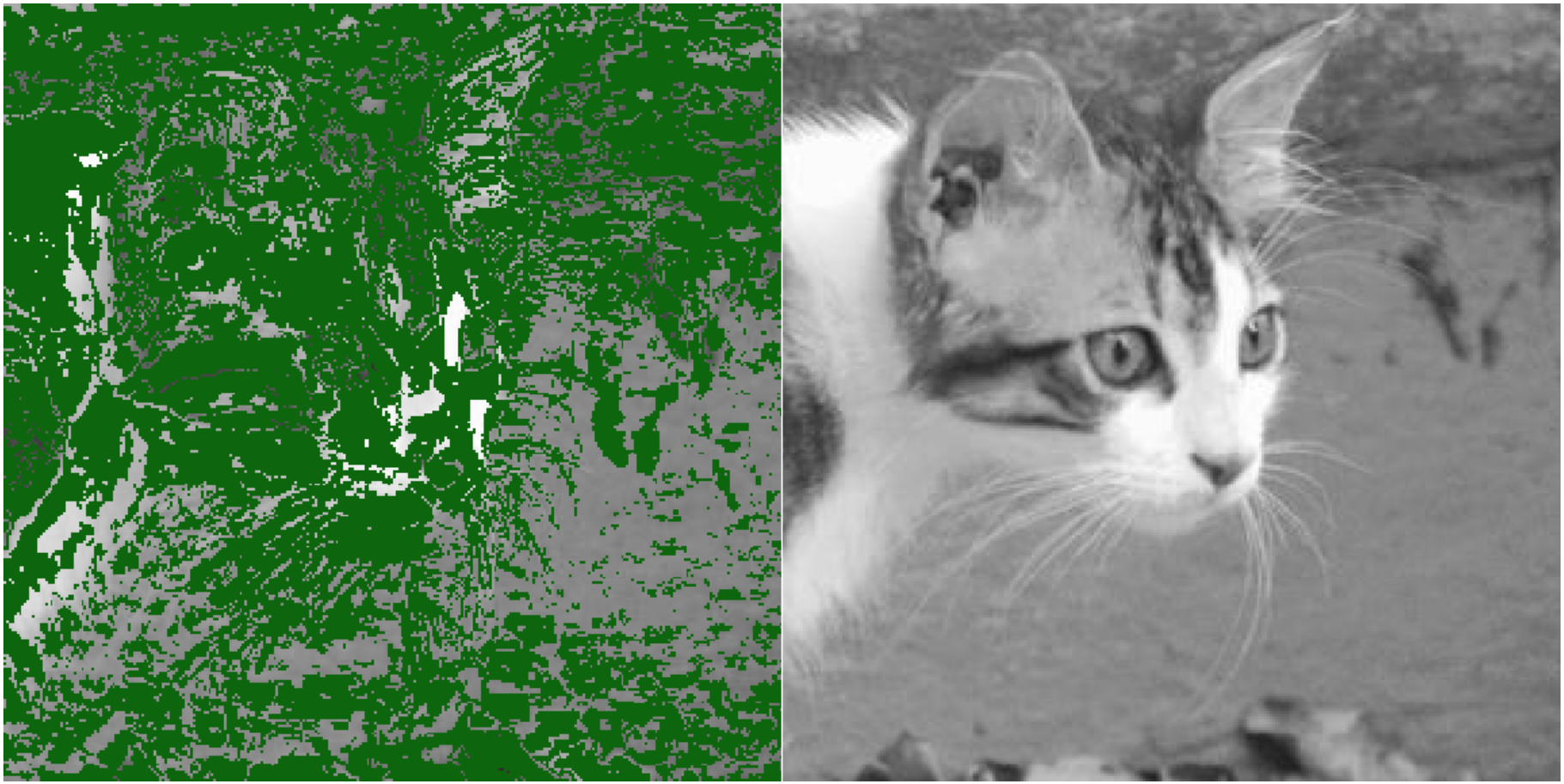


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

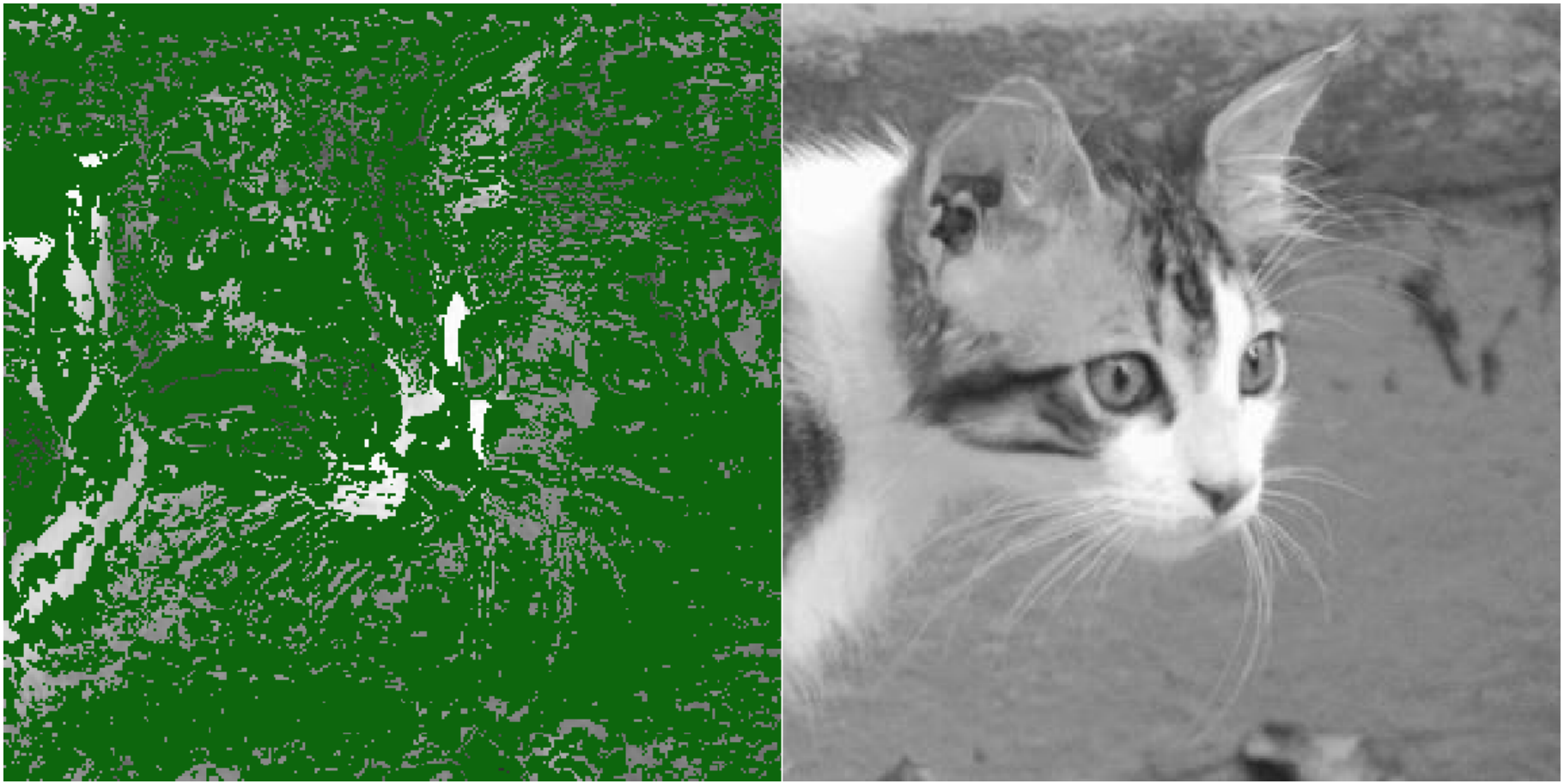


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

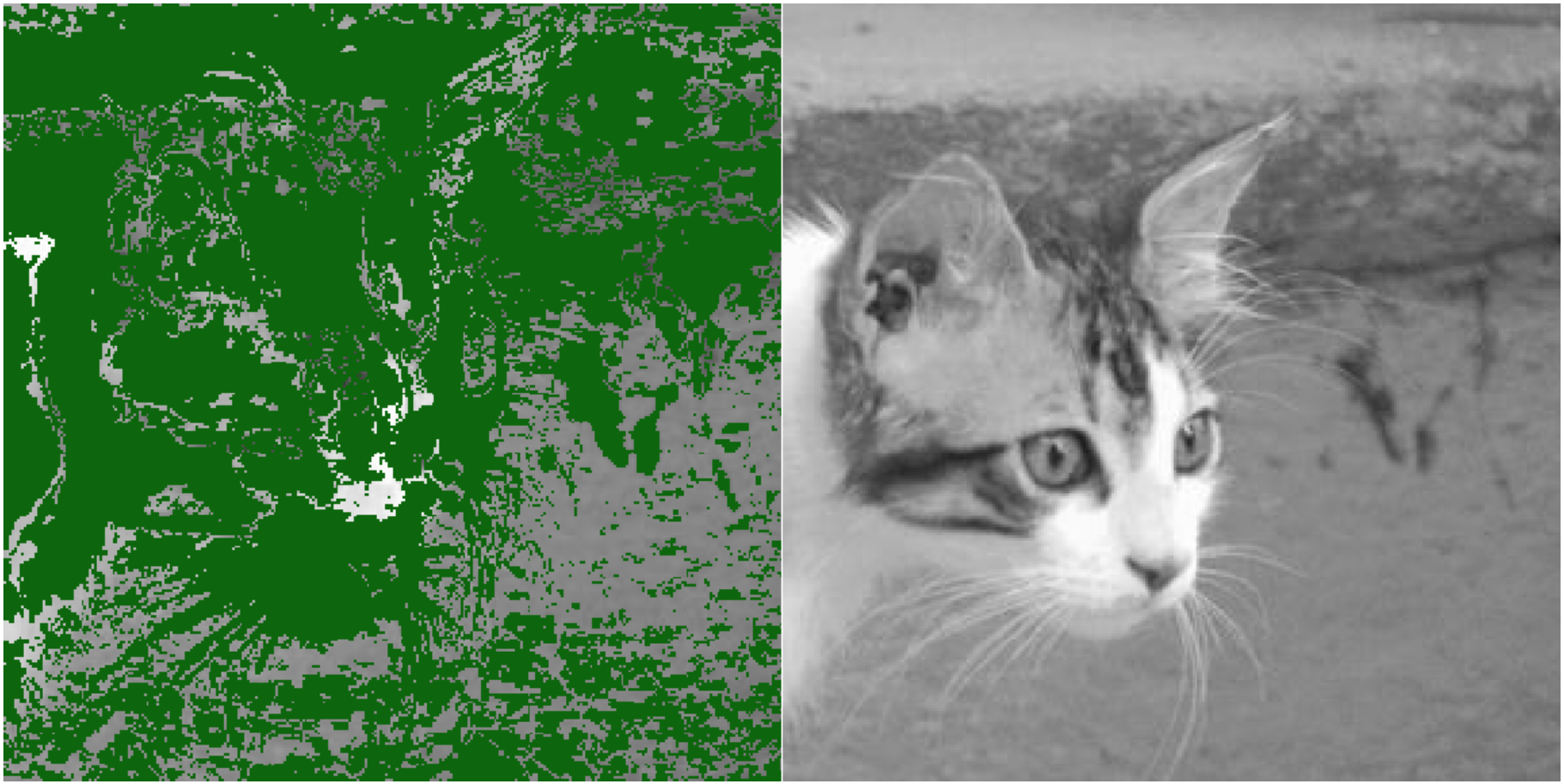


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

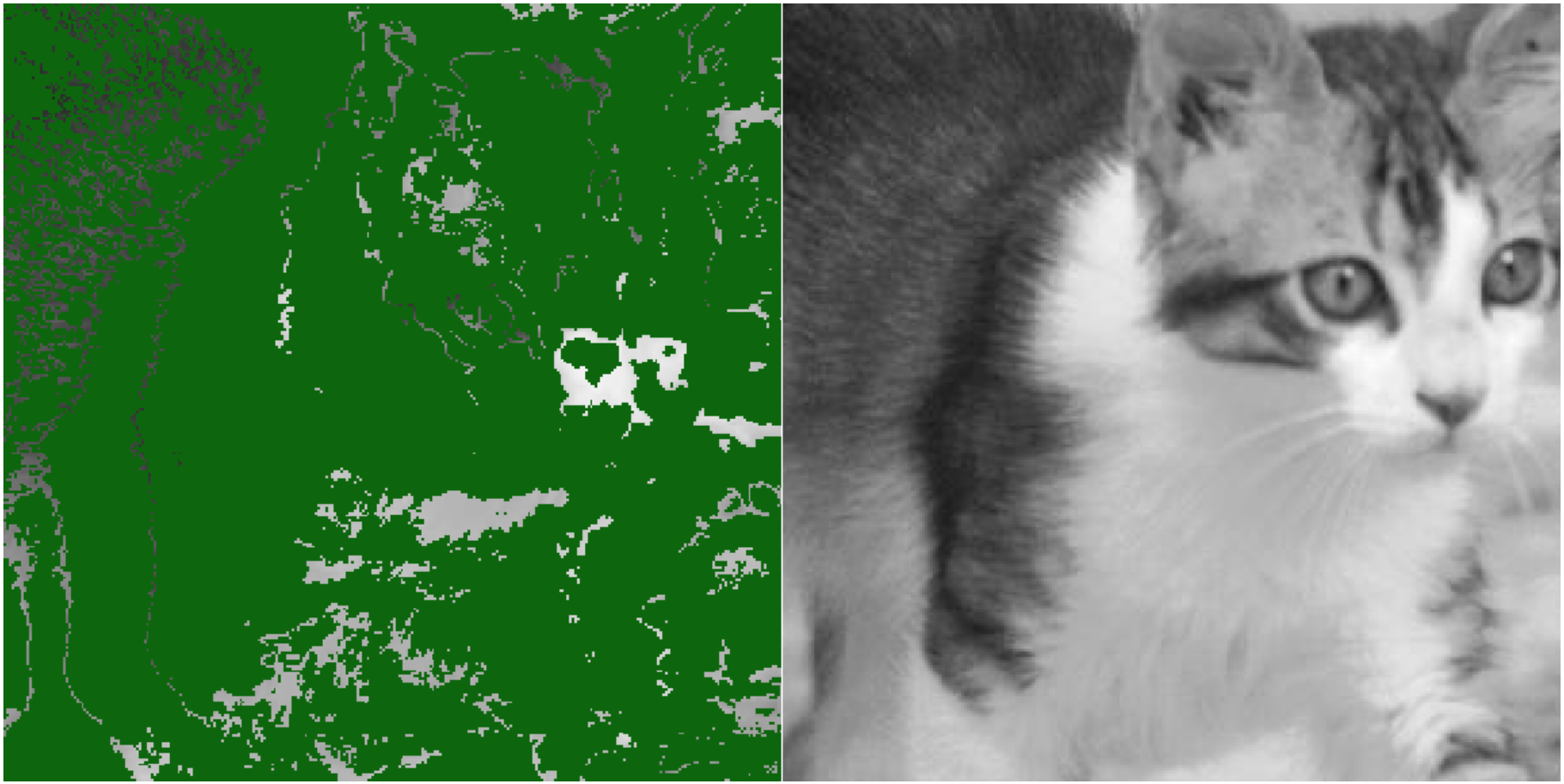


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

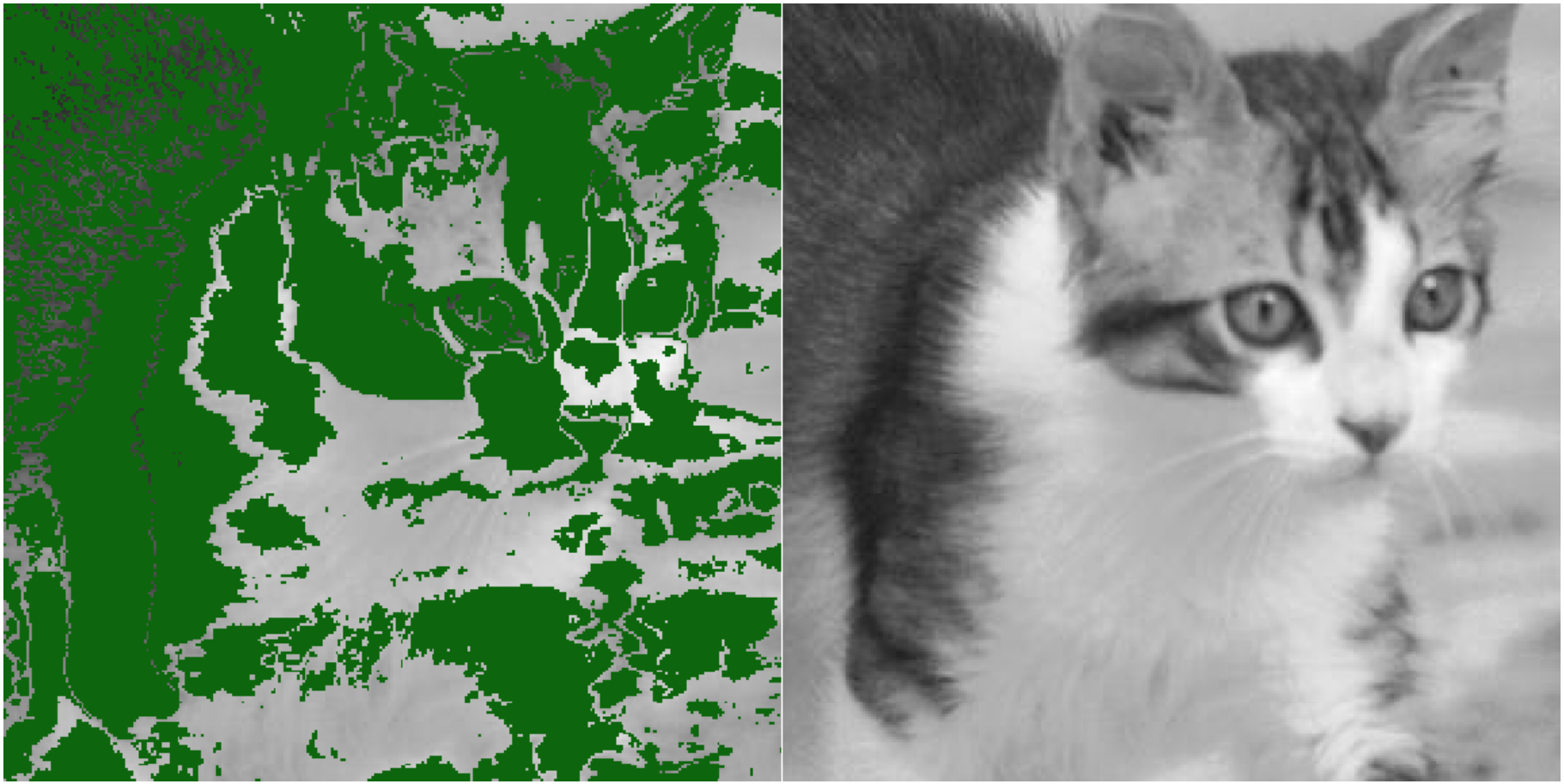


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying translations

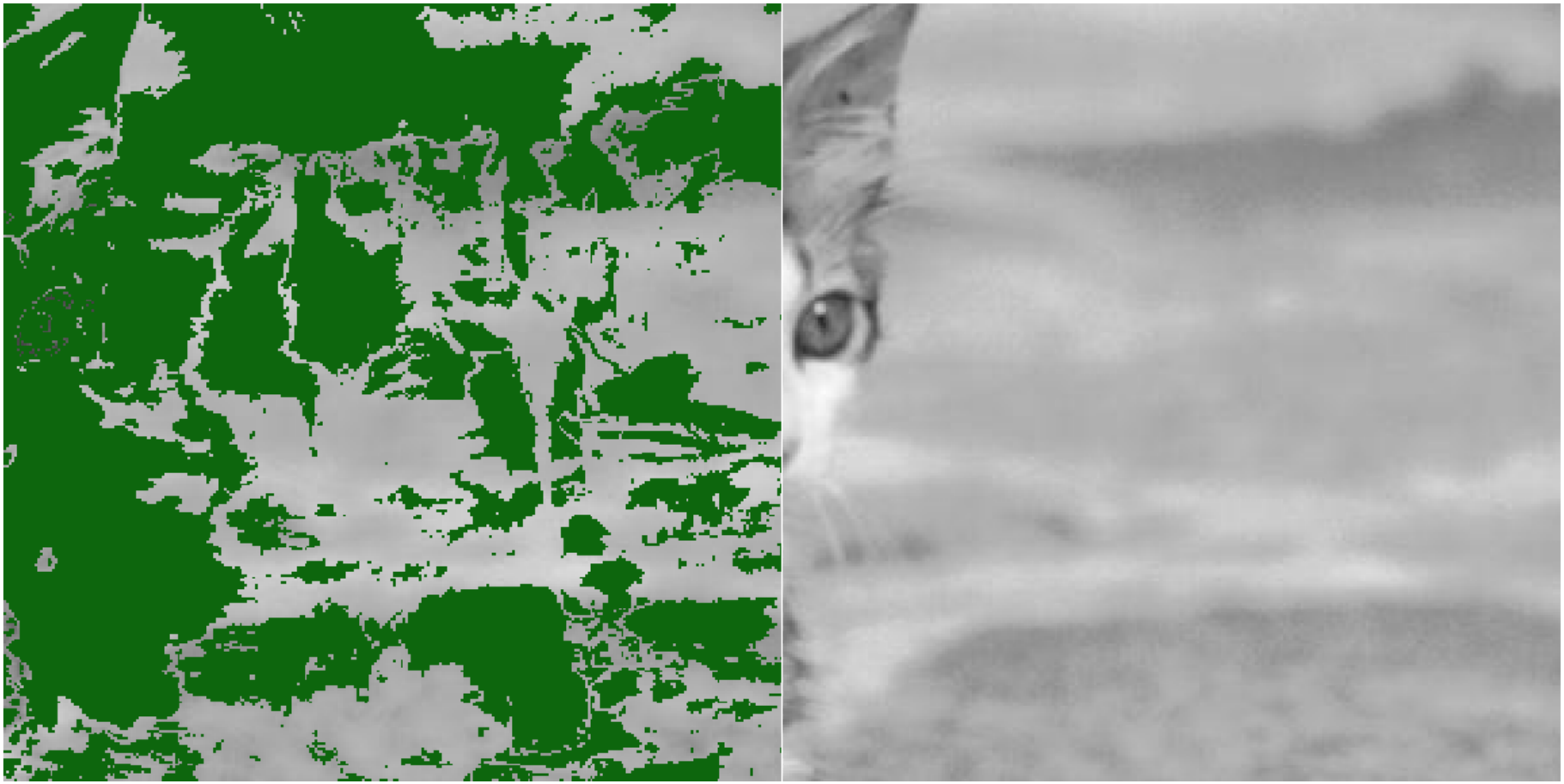


Figure: Visualization of a syntax label detecting some kind of translation.

Identifying related patterns coming from different images/scenes

First identification of the Kitten's head

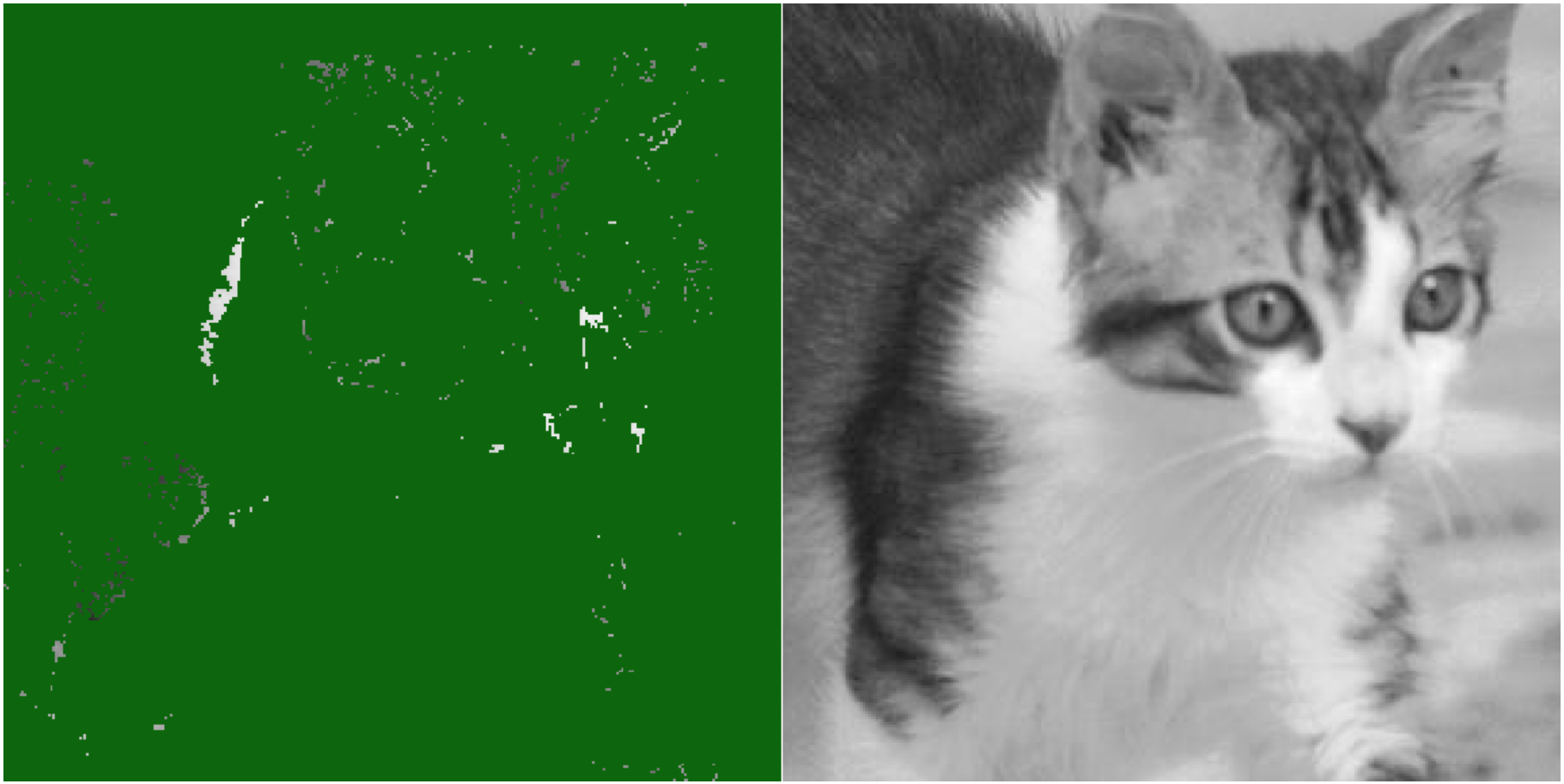


Figure: Visualization of a fixed syntax label recognizing related pattern.

Identifying related patterns coming from different images/scenes

First identification of the Kitten's head

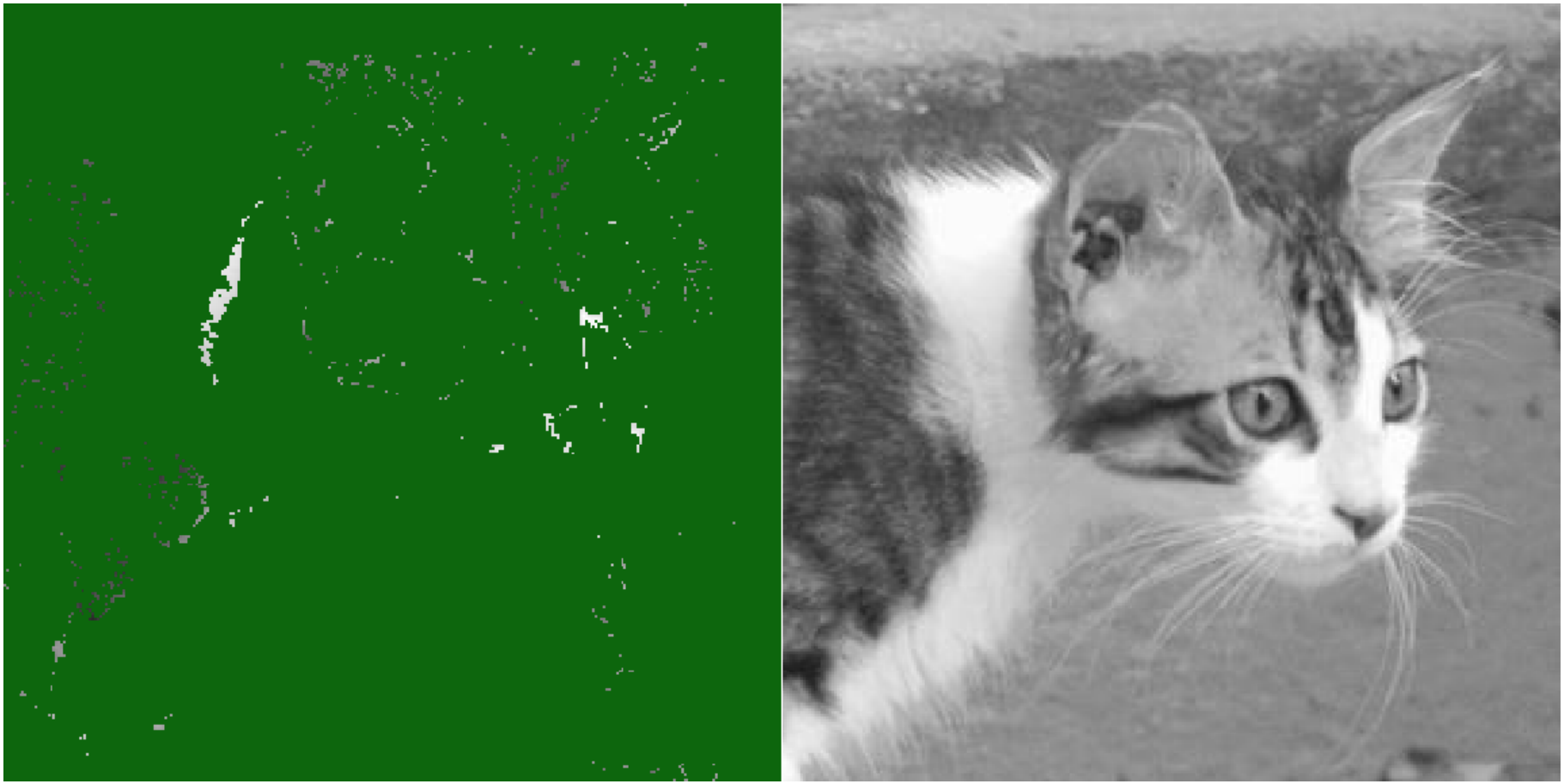


Figure: Visualization of a fixed syntax label recognizing related pattern.

Second identification of the Kitten's head

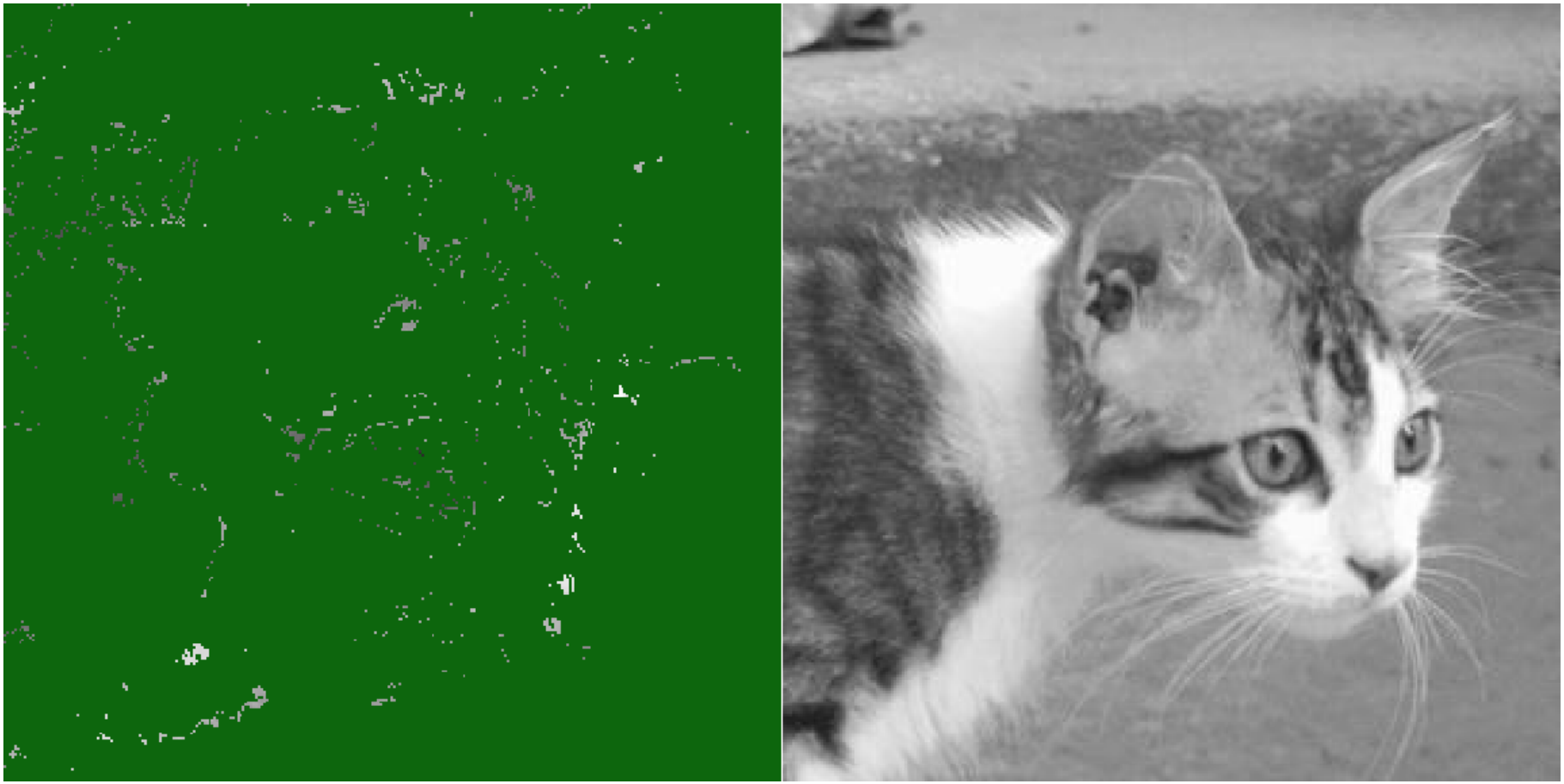


Figure: Visualization of a fixed syntax label recognizing related patterns.

Second identification of the Kitten's head

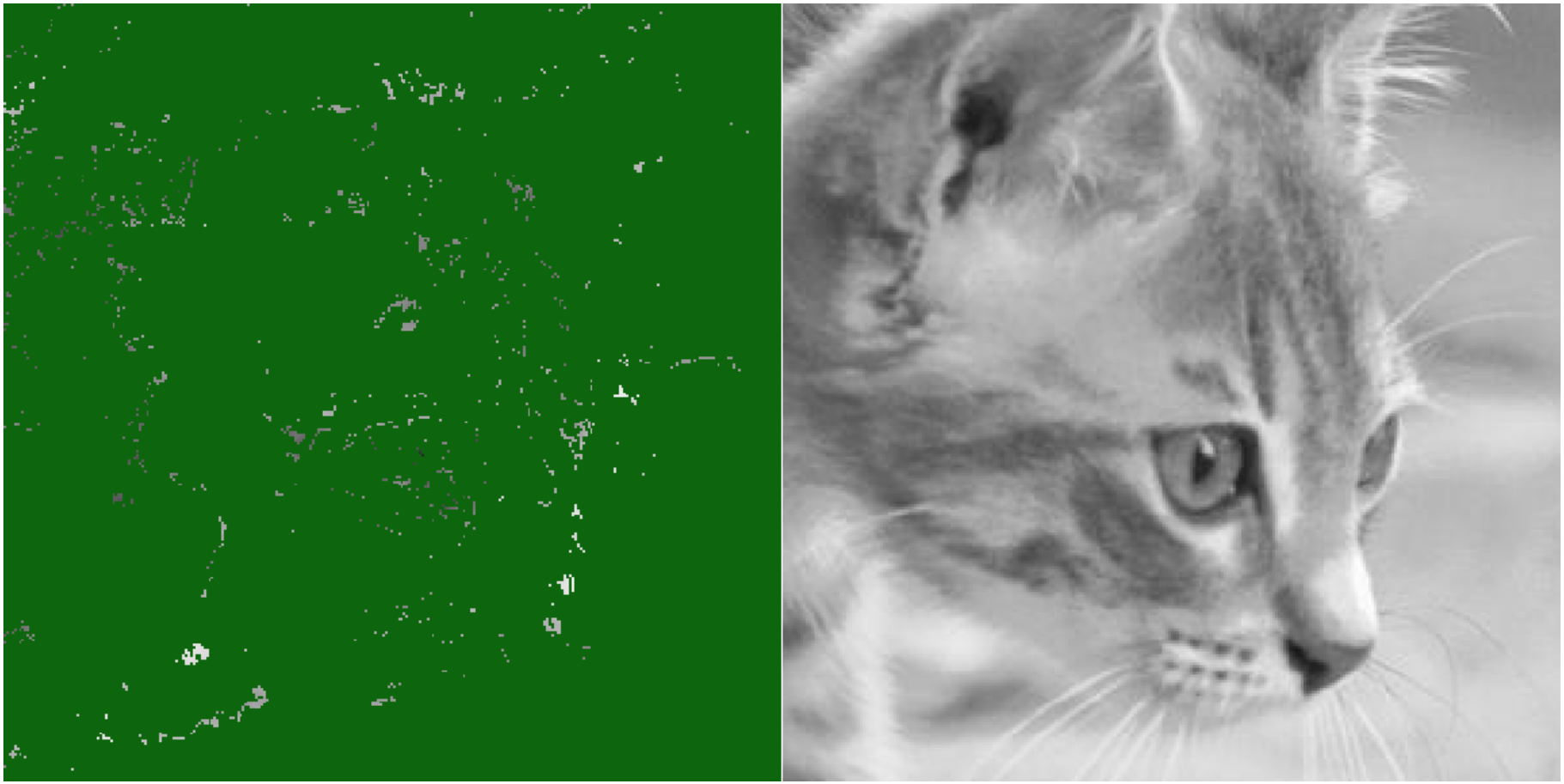


Figure: Visualization of a fixed syntax label recognizing related patterns.

Weak distortion and information fragmentation

- To get a weaker notion of distortion, introduce a probabilistic description of X as $\mathbb{P}_{S,V|X}$.
- Define the triplet (X,S,V), where $\mathbb{P}_{S|X} = \mathbb{P}_S = \frac{1}{d}\sum_{s=1}^{d}\delta_s$ is uniform on pixel locations and where $\mathbb{P}_{V|X,S} = \mathcal{N}\big(X_S,\sigma^2\big)$ is the value X_S at pixel S blurred by a Gaussian noise.
- Describe the sample by $\overline{\mathbb{P}}_{X,S,V} = \overline{\mathbb{P}}_X \mathbb{P}_{S,V|X}$.
- Introduce the empirical weak distortion

$$\widetilde{\mathcal{D}}(\overline{X},B,C) = 2\sigma^2\Big[1-\exp\Big(-\inf_{Q\in\mathcal{Q}}\mathcal{K}\big(Q_{X,S,V},\overline{\mathbb{P}}_{X,S,V}\big)\Big)\Big]$$

$$= 2\sigma^2\overline{\mathbb{P}}_X\Big[\min_{A\in\mathcal{T}_{B,K}}\mathbb{P}_S\sum_{j\in A}\mathbb{1}\big(S\in B_j\big)\big[1-\exp\big(-(X_S-C_{j,S})^2/(2\sigma^2)\big)\big]\Big]$$

$$\leq \mathcal{D}(\overline{X},B,C), \text{ where}$$

$$\mathcal{Q} = \Big\{Q_{X,S,V} : \text{for some } A:\mathbb{R}^d\to\mathcal{T}_{B,K},\ Q_{V|X,S} = \mathcal{N}\big(C_{A(X),S},\sigma^2\big)\Big\},$$

and its expected value $\widetilde{\mathcal{D}}(B,C)$.

Lemma

Minimizing the weak distortion in C, the fragment contents, gives the optimal approximation of $\overline{\mathbb{P}}_{X,S,V}$ under a conditional independence assumption only. Indeed

$$\inf_{C_j \in \mathbb{R}^{B_j}} \widetilde{\mathcal{D}}(\overline{X}, B, C) = 2\sigma^2 \inf_{Q \in \mathcal{Q}} \left[1 - \exp\left(-\mathcal{K}\left(Q_{X,S,V}, \overline{\mathbb{P}}_{X,S,V}\right)\right)\right]$$

where

$$\mathcal{Q} = \left\{ Q_{X,S,V} : \text{ for some } A : \mathbb{R}^d \to \mathcal{T}_{B,K}, Q_{V|X,S} = Q_{V|S,\ell_{A,B}(X,S)} \right\}$$

and

$$\ell_{A,B}(X,S) = j \iff S \in B_j \text{ and } j \in A(X)$$

is the classification function defined by A and B.

Proposition

Let $(\widehat{B},\widehat{C}) \in \arg\min_{(B,C)\in\mathcal{M}(\mathcal{S})} \widetilde{\mathcal{D}}(\overline{X},B,C) + 2\sigma^2\left(\frac{\log(n\mathcal{S}^{-1}K^{-1})}{\log(2)}\sqrt{\frac{8\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}}\right.$

$$\left. + 2\sqrt{\frac{\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}}\right) + 2\sigma^2\sqrt{\frac{(\sqrt{2}+1)\big(1+2\log(e|\mathcal{T}_{B,K}|)\big)\mathcal{S}K}{n}},$$

with probability at least $1-\delta$,

$$\widetilde{\mathcal{D}}(\widehat{B},\widehat{C}) \leq \inf_{(B,C)\in\mathcal{M}(\mathcal{S})} \widetilde{\mathcal{D}}(B,C) + 2\sigma^2\left(\frac{\log(n\mathcal{S}^{-1}K^{-1})}{\log(2)}\sqrt{\frac{8\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}}\right.$$

$$\left. + 2\sqrt{\frac{\mathcal{S}K\log(|\mathcal{T}_{B,K}|)}{n}}\right) + 2\sigma^2\sqrt{\frac{(\sqrt{2}+1)\big(1+2\log(e|\mathcal{T}_{B,K}|)\big)\mathcal{S}K}{n}}$$

$$+ 2\sigma^2\sqrt{\frac{k(k-1)\log(2)+2\log(\delta^{-1})}{n}}.$$