Algorithm Acceleration of Compressed Sensing with Cloud

Zhang Yong-Ping, Zhang Gong-Xuan and Zhu Zhao-Meng
School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China

Abstract: Compressed Sensing (CS) is a new sampling approach. It can accomplish simultaneously sampling and compressing. But the signal is reconstructed by solving a numerical optimization algorithm with high computational complexity in the theory of CS. The longer time of reconstruction is disadvantageous to the applications of CS. To overcome this problem, a complete scheme named Cloud_CS is introduced in this study. The scheme can speed up CS algorithms written in Python. Cloud_CS is built on OpenStack cloud computing platform. It achieve algorithms acceleration by serializing function, rewriting the function pmap and defining two marks which are named #remote and #parallelize. The experiments show that the reconstructed effects of signals is substantially equivalent to running locally and the running time can be reduced significantly. Then, Cloud_CS scheme is correctness and effectiveness.

Key words: Compressed sensing, cloud, algorithm acceleration, code migration, parallelization algorithm

INTRODUCTION

In recent years, a new approach of signal acquisition named Compressed Sensing (CS) is attracting more and more attention in signal processing, information science, electrical engineering, statistics and computer science after the theory was put forward in 2006 (Eldar and Kutyniok, 2012). It can represent a sparse or compressible signal using a very small number of measurements in some bases and has the ability to reconstruct the original signal with high accuracy by solving a numerical optimization algorithm (Boufounos et al., 2007) where the traditional methods, such as Nycquist rate sampling theorem, have failed. In other word, compressed sensing can directly acquire the compressed form by throw away the redundant information when the signal is sampled. Thus, the amount of collected data is relatively small and the full-length signal can be reconstructed from the small sampling. This characteristic presents a good prospect of applicability in the field of information acquisition.

But compressed sensing is an algorithm with high computational complexity. For example, the calculation scale of reconstruction is equivalent to solving a linear programming problem with 8192×262144 when the length of signal is 8192 in BP algorithm (that is one of the compressed sensing algorithms (Chen et al., 2001). So, the acceleration of CS algorithm is essential. Cloud computing technology (Armbist et al., 2010) can provide speedup results in some case because it can utilize the existing computing resources in the networks and avoid one-off investment. Saving investment makes it very attractive, especially for private physical networks.

An acceleration scheme of CS algorithm based on cloud technology, named Cloud_CS, is described in this article. Cloud_CS can migrate automatically the complex reconstructed operations to the cloud environment, so it can get more computing resources provided by cloud to speed up the algorithm.

COMPRESSED SENSING

Compressed sensing (Donoho, 2006), also known as compressive sampling (Candes, 2006) or compressive sensing (Baraniuk, 2007) is a new sampling approach to simultaneous sensing and compression. It can potentially reduce the sampling and computation costs for acquisition of signals which have a sparse or compressible representation in some basis. In this section, a brief description of compressive sensing theory will be presented (Tsai and Donoho, 2006; Candes and Wakin, 2008).

In general, the compressible original signal X, not necessarily sparse, has the sparse representation Θ based on a transformation basis \( \Psi_{sys} \):

Corresponding Author: Zhang Yong-Ping, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

269
where, conversion coefficient \( \Theta \) is sparse that the vast majority entries in matrix \( \Theta \) are zero (Elad, 2010). Then, \( \Theta \) can be compressed by using measurement matrix \( \Psi_{M \times N} \) \( (M \ll N) \). The compressed data \( Y \) is defined in Eq. 2:

\[
Y = \Phi \Theta
\]

where, the specially designed \( \Phi \) must be independent of the transformation basis \( \Psi \). The compression ratio here is \( N : M \). Generally, the compression ratio can become bigger if coefficient \( \Theta \) is sparser.

In order to reconstruct original signal \( X \) from the compressed data \( Y \), according to the compressed sensing theory, the following \( \ell_0 \)-norm optimization problem need to be solved:

\[
\min |\Theta|_0 \text{ s.t. } Y = \Phi \Theta
\]

where, \( |\Theta|_0 \) is the number of nonzero entries in \( \Theta \).

Equation 3 is a nonlinear optimization problem and cannot be solved because it is a \( \text{NP}-\text{hard} \) (Dhariwal, 2007) problem. However, Chen et al. (2001) have proposed that a \( \ell_1 \)-norm optimization problem shown at Eq. 4 can substitute for the equation 3 when the \( \Phi \Psi^T \) obey the Restricted Isometry Property (RIP) (Foucart and Lai, 2009) theorem:

\[
\min |\Theta|_1 \text{ s.t. } Y = \Phi \Theta
\]

Equation 4 is a convex optimization problem and it can be easily solved. The necessary and sufficient condition of RIP is that the measurement matrix \( \Phi \) must be independent of the transformation basis \( \Psi \).

The computing model of compressed sensing theory is shown in Fig. 1.

**DESIGN OF ACCELERATION SCHEME**

**Design:** Usually, the algorithm acceleration is running programs by using more or better computing resources, such as more CPUs, more memories and faster hardware. The cloud can be convenient to provide a large number of computing resources for the user according to demand by networks. OpenStack (2012) is one of the most popular frameworks for running applications on cloud because it is simple to implement and massively scalable. OpenStack is an Infrastructure as a Service (IaaS; Bharadwaj et al., 2010) cloud computing project. It consists mainly of two modules: A cloud computing fabric controller-Nova and a massively scalable redundant storage system-Swift.

![Fig. 1: Theoretical framework of compressed sensing](image)

![Fig. 2: Theoretical execution process of acceleration scheme](image)
Cloud CS can use the computing resources of cloud platform to speed up CS algorithm written in Python (Cai et al., 2005) and provide on-demand services for users. In general, the collected data from a signal are acquired, transmitted and stored in compressed form in accordance with this scheme. The complex reconstruction operation will be migrated automatically to the cloud environment when users need to reconstruct the original signal. And then, cloud allocates some computing resources based on the needs of the users to speed up the reconstruction algorithm and eventually returns the calculated results.

The design scheme of Cloud CS focuses on algorithm migration and parallelization:

Algorithm migration: How to migrate the code to cloud? It is based on functions in Cloud CS scheme. In the scheme, a mark #remote is defined. If the mark #remote appears in the front of a function, the function will be migrated to cloud when it is called. For example, the presentation:

```python
#remote
def reCloud (...):
```

indicates that the function reCloud will be migrated to cloud platform.

Generally, there are two ways to achieve the function migration. (1) To package the entire application and send it to cloud. This method is simple to implement, but it will send more redundant information and increase the time of data transmission. (2) To serialize function objects and variables, then package the function and send it to cloud. The serialization process involves serializing function, operating environment and results. This method can transmit the minimum amount of data, but it is complex and need to study the language specific features.

The second method is chosen in the design of Cloud CS. The python objects are serialized by rewriting two functions. The functions are dumps function to serialize and loads function to de-serialize. They are based on the built-in modules of marshal (http://docs.python.org/2/library/marshal.html) and pickle (http://docs.python.org/lib/module-pickle.html). At the same time, another module named “Flasky” is defined in the scheme. It can serialize more python objects, include functions, classes and running environment. The procedure of function migration can be shown in Fig. 3.

![Diagram](image)

Fig. 3: Running procedure of a function with the mark #remote when it is called. The procedure includes function call, collecting information, serializing and de-serializing, running, return results. The related informations of function collected include variables, input and output, environmental parameters, the results, etc.

another mark #parallelize is defined in the scheme. The codes will be processed in parallel if they are marked by #parallelize. For example, the presentation:

```python
#parallelize
for i in --:
```

indicates the for statement will be processed in parallel. The loops are mainly for statement, while statement, list comprehensions, set comprehensions and dict comprehensions in Python. Because the CS algorithm mainly uses for statements, look at the parallelization of for statement, for example.

In Python, the list comprehensions can provide a concise way to create lists. Some for statements also can be transformed into list comprehensions. For example, there is a loop of for:

```python
[(a, b) for a in [2, 3, 4] for b in [1, 3, 4] if a = b]
```

The result is:

```plaintext
[(2, 1), (2, 3), (2, 4), (3, 1), (3, 4), (4, 1), (4, 3)]
```
So, the loop of for is equivalent to the list comprehension:

```python
combs = []
for a in [2, 3, 4]:
    for b in [1, 3, 4]:
        if a != b:
            combs.append((a, b))
combs
```

Then, the for statement can be transformed into a list which can easily achieve the parallelization.

In addition, the "map", a built-in function in Python, is used to create recursive procedure. The function is not suitable for parallelization. Its migration and parallelization is achieved by rewriting it to pmap function. A new class that inherits the list comprehensions is defined in pmap function. It can not wait for all the results to continue to perform the following iteration by using the new class in the iteration operations and then parallel processing can be achieved.

Most of the design can be supported by famous scientific packages in Python, such as scipy (Jones et al., 2001), numpy (Developers, 2010) and matplotlib (Hunter, 2007).

**ALGORITHM AND ITS MODIFICATION**

Signal reconstruction is achieved by solving numerical optimization problems in compressed sensing. It has different reconstruction algorithms depending on different optimization methods. Currently, the popular reconstruction algorithms of compressed sensing can be grouped into three broad categories: Convex relaxation method, greedy algorithm and combination algorithm. Basis Pursuit (BP) algorithm is a representative algorithm of convex relaxation method. Greedy algorithm mainly includes Matching Pursuit (MP), Neff and Zakhor, 1997 algorithm and its improved algorithm.

**OMP algorithm:** As one of the greedy algorithms, OMP algorithm (Tropp and Gilbert, 2007) is a sparse approximation algorithm. The major advantages of this algorithm are easy to implement and its speed. Its computational complexity is lower than BP algorithm and its iteration times are less than MP algorithm. The experimental verification of this study is to use OMP algorithm.

The main idea of the OMP algorithm is to choose the column of measurement matrix \( \Psi \), at each iteration, which is related to the greatest extent with the present redundant vector by solving the following optimization problem:

\[
\lambda_i = \max_{\lambda_i \neq \lambda_j \neq \ldots \neq \lambda_M} |x_i^T \Psi| \tag{5}
\]

where, \( \lambda_i \) is the index of column at t-th iteration, \( r_{\lambda_i} \) is the residual and \( q_i \) is the t-th column of measurement matrix \( \Phi \). Then a new estimated signal can be obtained by solving the least-square method:

\[
x_t = \min_{x} |\Phi x - y|_2 \tag{6}
\]

where, \( x_t \) is the solutions of t-th K-sparse approximation, followed by subtracting the related part from measurement vector and iterate on the residual.

The pseudo code of OMP algorithm can be shown in the following algorithm description (Tropp and Gilbert, 2007) and the flow chart is shown in Fig. 4.

**Input:**
- Measurement matrix \( \Phi \)
- Sampling vector \( y \)
- Sparse degree \( K \)

**Output:** The K-sparse approximation solution \( x^* \) of original signal \( x \).

**Restructuring procedure:**
- Initialization: \( r_0 = y, \lambda_0 = \emptyset, t = 1 \)
- To obtain \( \lambda_t \) by solving Eq. 5
- Update \( \lambda_t = \lambda_{t-1} \cup \{\lambda_t\} \) and \( \Phi = [\Phi_{\lambda_t}, \Phi_\emptyset] \)
- To obtain new approximation solution \( x_t \) by solving Eq. 6
- Update \( r_t = y - \mu_t x_t \), \( t = t+1 \)
- IF (t<K THEN return 2)

**Algorithm migration:** Using OMP algorithm, the original signal can be reconstructed or solved an approximation solution. In order to migrate the algorithm to cloud platform, the mark of #remote must be added in front of the function. That is:

```python
#remote
def OMP(y, \Phi, k):
    function-body
```

Here, OMP(y, \( \Phi \), k) is the name of OMP algorithm. It has been previously defined. With the above definition, the OMP function can be automatically serialized and sent to the cloud environment in each calls.

**Parallelization:** The OMP algorithm can reconstruct the signal from a vector \( X_{\text{real}} \). It need to repeatedly perform OMP algorithm to reconstruct a signal when the signal \( X \) is \( N \times M \). This can be a substantial increase in execution time of reconstruction algorithm. The design scheme is mainly devoted to this case. This scheme can run multiple OMP algorithms at the same time and send them to
Initialization
\( R_0 = y, \ V_0 = \emptyset, t = 1 \)

Choose the index \( i_t \) by solving \( \max_{N \in I, \ N \neq 1} |<r_{t-1}, n_i>| \)

Renew the index set \( V_t = V_{t-1} \cup \{i_t\} \) and the sensing matrix \( N_t = (M_{2t}, r_{t0}) \)

 Obtained a new signal estimate by solving \( \min_x \|N_t x - y\|_2 \)

Renew the residual \( r_t = y - N_t x_t, t = t + 1 \)

Get the estimate of the ideal signal

Yes

\( t < M \)

No

small pieces with the same length when \( N \) is larger and then the signal \( X_{N_t} \) may be transformed into the other signal \( X_{N \times N_t} \), where the \( N1 \) is the length of a small piece and the \( N \) is equal to \( N1 \times N2 \) or the \( N+1 \) is equal to \( N1 + N2 \) \((s < N1)\). So, the scheme can also be used to speed up the reconstruction algorithm.

For high-dimensional signals, they can also be converted to the forms of signal \( X_{r_{PL}} \). For example, consider the signal \( X_{N^D \times N^T} \). It can be transformed into the signal \( X_{N^D \times N^T} \), i.e., \( X_{N \times L} \) where \( L \) is equal to \( S \times T \). The signals \( X_{N^D \times N^T} \), \( X_{N \times L} \) and other high-dimensional signals can be taken the same treatment. Then, the scheme can also be used.

The parallelization of loops also requires the mark \#parallelize. For example:

```markdown
#parallelize
for i in range(N):
X[:, i] = OMP(Y[:, i], reshape(C, 1, 1), Φ, k)
```

Like the above for statement, the loops will be automatically executed in parallel if \#parallelize is marked before the loops statement.

**EXPERIMENTS AND ANALYSIS**

**Experimental environment:** In the experiments of this study, the local computing platform for running test programs is a personal computer with 2.4 GHz Intel(R) Core(TM) 2 Duo dual-core processor (E4600), 2 GB memory, 64 bits, 10/100 Mbps wired Ethernet and running windows 7 SP1.

An OpenStack IaaS platform has been built on a cluster of IBM BladeCenter rack servers. For each server, there is a 2 GHz Intel Xeon quad-core processor, 24 GB memory, 64 bits, 1000 Mbps wired Ethernet and runs Ubuntu server edition. A various numbers of tiny VM instances (1VCPU/0GB Disk/512MB Ram) will be launched for these experiments in the cloud computing environment. For computing resource is defined as a tiny VM instance in these experiments.

Next, this scheme will be verified via well-designed three groups of experiments. These three groups of experiments are used to verify the correctness of the algorithm, accelerating effects with different numbers of computing resources and accelerating effects for the different sizes of signals. For each group of experiments, the reconstruction algorithm should be repeatedly executed and then calculate its average running time.

**Verification of correctness:** This group of experiments will test the impact on reconstruction results when the

---

Fig. 4: Flow chart of OMP algorithm. It can approach the original signal by gradually iterative approximation
different computing resources to perform simultaneously, thereby improving execution speed of the reconstruction algorithm.

For the signal \( X_{N \times N_t} \), the reconstructed time is smaller too and it does not require using any accelerated program when \( N \) is smaller. The signal can be divided into multiple
algorithm is migrated and parallelized. It will reconstruct a standard image named lena.bmp with 256*256 pixels on the local platform and in the cloud computing environment. These reconstruction effects are shown in Table 1 and Fig. 5.

Consider the comparison of PSNR (Power Signal-to-Noise Ratio) in Table 1. The PSNR of reconstructed signal is almost the same whether the codes are run in the local or sent to cloud, the nuances exist only because there is difference in the sparse transform and observation matrix. As Fig. 6 shows, the approximating value of original signal can be reconstructed and accepted.

**Effects of different numbers of computing resources:**
This group of experiments will observe the effects of algorithm acceleration when the number of computing resources varies. Similarly, this group of experiments will reconstruct the standard image named lena.bmp with 256*256 pixels. It changes the number of computing resources from 1 to 16. For various numbers of computing resources, the average execution time and speedup are computed after the reconstruction algorithm is repeatedly executed. The results of this group of experiments can be shown in Fig. 6 and 7.
Table 2: Comparison of speedup effects when the sizes of signal are different. The first column is the size of reconstruction signal. For every signals, the No. of computing resources used to test are 1, 8, and 16.

<table>
<thead>
<tr>
<th>Image size</th>
<th>No. of computing resources</th>
<th>Average execution time (sec)</th>
<th>Speedup</th>
<th>Speedup/No. of resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>64*64</td>
<td>1</td>
<td>0.9522</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.3710</td>
<td>3.42</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.1911</td>
<td>4.85</td>
<td>0.30</td>
</tr>
<tr>
<td>128*128</td>
<td>1</td>
<td>5.9534</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.1390</td>
<td>5.23</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.7085</td>
<td>7.54</td>
<td>0.47</td>
</tr>
<tr>
<td>256*256</td>
<td>1</td>
<td>47.5001</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6.6415</td>
<td>7.14</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>5.7026</td>
<td>12.57</td>
<td>0.79</td>
</tr>
<tr>
<td>512*512</td>
<td>1</td>
<td>455.1829</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>60.5982</td>
<td>7.51</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>31.3774</td>
<td>14.51</td>
<td>0.91</td>
</tr>
<tr>
<td>1024*1024</td>
<td>1</td>
<td>6056.9560</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>794.6400</td>
<td>7.62</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>400.8561</td>
<td>15.11</td>
<td>0.94</td>
</tr>
</tbody>
</table>

CONCLUSION AND FUTURE WORK

In this study, an algorithm acceleration framework based on the cloud computing platform is proposed and implemented. The framework can speed up the program written in Python (not limited to CS algorithms, although it is developed for compressed sensing). The accelerated scheme can achieve algorithm acceleration by defining two marks #remote and #parallelize and rewriting the function pmap. The mark #remote can send automatically the function to cloud environment, the mark #parallelize can execute the loop in parallel and the function pmap can break the order of iteration. The experimental results prove the framework's correctness and effectiveness. For example, the execution time is decreased from 47 to 4 sec by using 16 computing resources in Fig. 6. The research of scheme is significant because it can solve the problem of high time complexity of traditional compressed sensing algorithms and broaden its applications.

Next work: The future researches will be focused on the following two aspects:

- Optimization of scheme and algorithm. In this study, the algorithm acceleration framework and test algorithm is achieved by our own. All codes are written in Python language and without any optimization. So the execution time is higher than Matlab program that can call the standard library. The future, these algorithms will be optimized in the experiment in order to improve the performance of the acceleration framework.
- To study the applications of compressed sensing in Internet of Things. The Internet of Things (Ashton, 2009; ITU, 2005), probably be also known as wireless sensor networks by somebody, is one the important technologies which can change humanity's life. It has become a national strategic emerging industry (Zhu et al., 2009; EU, 2009; ITS...
Headquarters, 2009; MIT, 2011). There are lots of collected data in large-scale Internet of Things. For example, a logistics monitoring system supported by Internet of Things in a supermarket, if you need to track the location and status information of 10 million products, the amount of data generated will reach 10 GB in one day and 3.65 TB in one year.

To overcome the problem, the application of compressed sensing is studied in Internet of Things because it can potentially reduce the amount of collected data. However, the high computational complexity of reconstruction algorithms will increase the costs of data processing. The Cloud_CS framework can have a good interaction with Internet of Things. The cloud computing platform can be deployed in Internet of Things because there are lots of computing resources and storage resources in Internet of Things, such as CPUS, memories, personal computers, servers, monitoring equipments and even GPGPU. Then, the acceleration framework can decrease the execution time of compressed sensing algorithm and Internet of Things can provide computing resources for the framework.

In addition, the OMP algorithm and some authentication algorithms are used in this experiment; some other reconstruction algorithm will also be explored.

ACKNOWLEDGMENTS

We wish thank my classmate Rui Zhi and Professor Xiaohui Liu of Brunel University for their valuable suggestions. The study’s work is supported by Jiangsu 973 project (No. BEK2011022), Natural Science Foundation of China (No. 61170035), Nature Science Foundation of China (No. 61272420) and NUST project (No. 2011YBXM18).

REFERENCES


