CE-004

EC 市場における中古品のユーザ需要に関するテキストマイニング解析
Text Mining Analysis of User Demand for Used Products in the EC Market

伊集院 大将‡　　石垣 綾‡
Hiromasa Ijuin　Aya Ishigaki

1. はじめに

消費者の手元に届く製品は、資源を加工することによって作られ、サプライチェーンを通じて消費者へと届けられる。資源の消費量は年々増加しており、近年では資源枯渇は世界的に深刻な問題の 1 つになっている[1]。国際連合環境計画(United Nations Environment Programme; UNEP)によると、資源の消費量は年々増え続けており、年間の資源消費量は、1970 年の 270 億トンから 2017 年には 920 億トンに達しているという調査結果が出ている[2]。また、資源の大量消費は、モノの大量廃棄や、資源枯渇を引き起こす。大量廃棄を軽減し、資源枯渇を防ぐための手段として、ユーザが使い終わった製品のリサイクルやリユースが挙げられる[3]。例えば、リサイクルは使用済みとなった製品から資源を取り出すことで、新たに資源を採掘する必要がなくなり資源枯渇を防ぐことができる。また、リユースは、使用済みとなった製品が廃棄されることなく、他の場所で使い続けられるようになり、大量廃棄の軽減につながる。したがって、中古品と販売するリユース市場の規模拡大は、資源枯渇を防ぐために期待されている不可欠である。

近年におけるリユース市場の多くは電子商取引(Electric Commerce; EC)で行われている。その市場規模は年々拡大を続けており、環境省によると、2020 年では 2.4 兆円に達していると報告されている[3]。EC 市場で取り扱われている製品は、ユーザが製品の購入について時間や場所を問わずに検討できるメリットがある。しかし、ユーザは写真や商品説明の文章だけで自身の望む性能や外観への需要を製品が満たしているか判断する必要があり、この商品説明とユーザ需要の不一致を起こすとユーザが不満を持つ原因となる。特に中古品については、例え同じ機種の製品であっても劣化や消耗具合によって性能に大きな違いが出るため、ユーザ需要を満たすことが難しいと考えられる。

本研究では、中古品に対するユーザ需要を明らかにするため、EC サイト上の中古品の購入者のレビューに対してテキストマイニングを行う。はじめに、購入者のレビューを形態素解析し、中古品に対してユーザ需要を示す単語を明らかにする。次に、判別分析を用いることでユーザのポジティブレビューとネガティブレビューを抽出する。最後に、単語に対して共起ネットワークを作成し、単語ごとの関連を示すとともに、ポジティブレビューとネガティブレビューで用いられている単語の違いについて考察する。

2. 研究手順

図 1 は、研究手順を表している。本研究では、EC サイトで販売されている中古ノート PC について、ユーザ需要を分析するため購入者レビューのテキストマイニング 2 つの Step により分析する。

Step 1 では、レビューデータ全体についてのテキストマイニングを行い、レビューデータ全体の傾向を調べる。はじめに Step1.1 で中古ノート PC の購入者レビューデータを抽出し、判別分析を用いてレビューデータ全体の概要について考察する。次に Step 1.2 では MeCab を用いた形態素解析を行い、レビューデータで使われている単語の名詞について着目し、*TF-IDF* の算出や共起ネットワークを生成する。

Step 2 では、評価ポイントごとのテキストマイニングを行い、ポジティブデータとネガティブレビューの違いについて分析する。まず、Step 1 から得られた中古ノート PC のレビューデータに対して Step.2.1 でポジティブ／ネガティブレビューデータそれぞれを抽出する。さらに、Step 2.2 で形態素解析を行い、ポジティブレビューでよく用いられる単語やネガティブレビューでよく用いられている単語(名詞)の違いについて考察する。

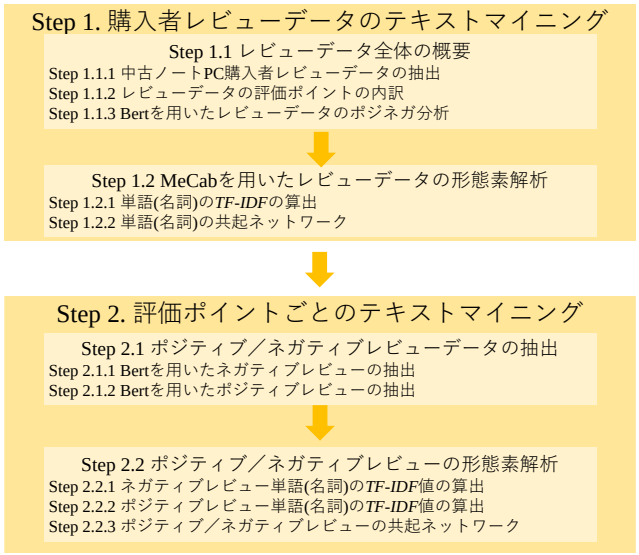


図 1.　全レビューデータにおける各単語の出現回数

3. 手法

本研究においては、MeCab を用いて形態素解析、BERT を用いた評判分析を行う。

BERT

BERT （ Bidirectional Encoder Representations from Transformers）とは、Transformer をベースとした自然言語処理の深層学習モデルである[4]。BERT は、汎用的な事前学習モデルを下層として使用し、各タスクに特化した上層のモデルをファインチューニングすることで、高い精度と効率を実現する。これにより、学習時間の短縮とモデルの性能向上を達成できる。本研究では、BERT による評判分析(Sentiment Analysis) [5]を行い、中古ノート PC に対するレビューの内容を評価する。

MeCab

MeCab とは、オープンソースソフトウェアにおける形態素解析エンジンの 1 つである[6]。本研究では、MeCab の API を用いることにより、形態素解析を行い、中古ノート

PC のレビュー内容について、文章から頻出している単語の名詞を抽出する。

TF-IDF

本研究において、レビュー内容における単語の重要度の指標として *TF-IDF* (Term Frequency-Inverse Document Frequency)を用いる[7], [8]。また、本研究においては、*TF-IDF* の算出に scikit-learn の API を用いていることより、式(1)によって求めることができる。

$$TF\text{-}IDF(t,d) = TF(t,d) \times (IDF(t,d) + 1) \quad (1)$$

TF(Term Frequency)とは、文書 d に対する単語 t の出現頻度$TF(t,d)$を意味し、式(2)によって求めることができる。

$$TF(t,d) = \frac{n_{t,d}}{n_d} \quad (2)$$

IDF(Inverse Document Frequency)とは、全文章に対して、単語 t を含む文書がどれくらいの頻度で存在するかを意味する逆文章頻度であり、scikit-learn の API を用いていることより、式(3)によって求めることができる。

$$IDF(t,d) = log\frac{1+\sum_d n_d}{1+tf(t,d)} \quad (2)$$

評判分析(Sentiment Analysis)

評判分析とは、対象に対する文章が筆者によって好意的(“POSITIVE”)、否定的(“NEGATIVE”)、または中立的(“NEUTRAL”)のいずれかのラベルであるかを判断する自然言語処理手法である[6]。また、それぞれのラベルにおけるラベル識別の確からしさを表す確信度(Confidence Value)は、ソフトマックス関数により式(4)と(5)によって得られる[9]。

$$f(x_i) = \frac{e^{x_i}}{\sum_i^I e^{x_i}} \qquad (i \in I) \quad (4)$$

$$f(x_i) = \frac{e^{x_i - x_{max}}}{\sum_{i=1}^I e^{x_i - x_{max}}} = \exp\left(x_i - x_{max} - \ln\left(\sum_i^I e^{x_i - x_{max}}\right)\right) \quad (5)$$

ソフトマックス関数とは、機械学習においてデータ i がラベルx_iである確率$f(x_i)$を式(4)で求め，式(5)に示す活性化関数で変換したものである[9]。

共起ネットワーク

共起ネットワークとは、文や単語の単位ごとに同時に生起する単語のペアを繋いでネットワークで可視化する分析法である[6]。本研究では、レビューデータの「レビュー内容」について、名詞の共起ネットワークを作成した。

4. 対象データの項目

本研究においては、ある EC サイトにおける購入者における中古ノート PC についてのレビューを対象とする。レビューデータについては、表 1 に示すデータ項目が含まれており、本研究で行う形態素解析や評判分析は、いずれかのデータ項目を対象に行った。

表 1. 対象のレビューデータ項目

No.	項目名	説明
1	商品名	販売されている商品の商品名(中古のだいたい製品には“中古”という単語も含まれている)
2	商品ジャンル ID	商品の種類によって一意に割り振られた数字
3	評価ポイント	購入した商品に対する購入者の商品の評価のポイント 最小値“1”、最大値“5”
4	レビュー内容	購入した商品に対する購入者のレビュー

5. 実験結果

5.1 購入者レビューデータのテキストマイニング

5.1.1 Step 1.1 レビューデータ全体の概要

Step 1.1 では、テキストマイニングを行い、対象としたレビューデータ全体の傾向を探る。はじめに、Step 1.1.1 でレビューデータから対象となる“中古ノート PC”のデータだけを抽出する。Step 1.1.2 と Step 1.1.3 では「評価ポイント」をもとに評判分析を行い、レビューデータ全体の傾向を調べる。

Step 1.1.1 中古ノート PC 購入者レビューデータの抽出

Step 1.1.1 では、EC 市場における中古ノート PC 購入者データの抽出を行った。具体的には、はじめにレビューデータにはノート PC 以外の商品についても、含まれているため、表 1 の「商品ジャンル ID」から、ノート PC のみを指定した。さらに、抽出したノート PC データには新品と中古品の両方が入っている。従って、表 1 の「商品名」に対して、“中古”という単語が入っている商品のみを抽出した。最後に、「レビュー内容」に対してテキストマイニングを行うことから、中古ノート PC のレビューデータの「レビュー内容」が空でないデータを抽出したところレビューの件数は 11,641 件になった。

また、購入者レビューデータの傾向を調べるために、抽出したデータのうち、表 1 に書かれている「評価ポイント」の値に対してそれぞれの割合を調べたところ、図 1 のような結果を得ることができた。図 1 より「評価ポイント」の項目については、“5”のデータの割合が最も多い 58%を閉めており、また評価ポイント“4”と合わせると全データの 90%を閉めている。このことから、今回分析を行ったレビューデータは全体的に好意的なデータを対象としていることがわかった。

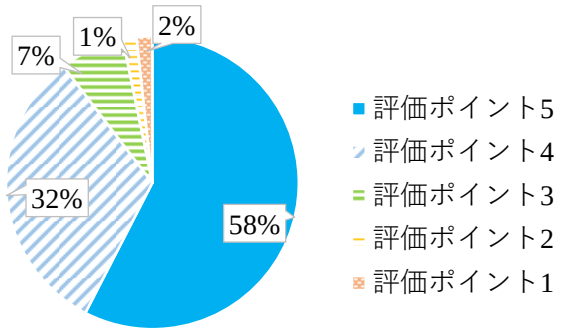


図 1. 全レビューデータにおける各単語の出現回数

Step 1.3 Bert を用いたレビューデータの評判分析結果

本項では、中古ノート PC 購入者のレビューにおける「レビュー内容」の項目に対して、Bert を用いて評判分析を行ったときの各評価ポイントにおけるポジティブレビューとネガティブレビューの割合について考察する。

図 2 は、各評価ポイントにおけるポジティブレビュー、ネガティブレビュー、ニュートラルレビューの割合を示したものである。「評価ポイント」の値が“1”や“2”のように購入者の需要を満たせていないものは、「レビュー内容」の評判分析もネガティブに分類された。反対に「評価ポイント」の値が“4”や“5”のように購入者の需要を満たしてい

るものは、「レビュー内容」の評判分析は、ポジティブに分類された。

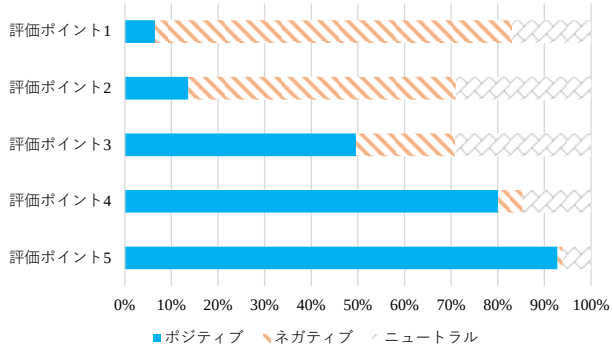


図 2. 評価ポイントごとのにおけるポジティブレビュー、ネガティブレビュー、ニュートラルレビューの割合

5.1.2 Step 1.2 MeCab を用いたレビューデータの形態素解析

Step 1.2 では、Step 1.1 で抽出した中古ノート PC の「レビュー内容」全体に対して、MeCab を用いた形態素解析を行う。Step 1.2.1 では、単語(名詞)ごとの *TF-IDF* を算出し、単語ごとの重要度を調べる。さらに、Step 1.2.1 で共起ネットワークから単語(名詞)ごとの関係性を調べる。

Step 1.2.1 単語(名詞)の *TF-IDF* の算出

Step 1.2.1 では、中古ノート PC 購入者のレビューにおける「レビュー内容」の項目に対して、MeCab を用いて名詞のみを抽出したときの形態素解析結果について述べる．図 3 は、全レビューデータにおける各単語(名詞)の TF-IDF を示している。図 3 から、“購入”という単語が最も大きく $TF\text{-}IDF^{購入} = 1{,}989$、次点の“満足”という単語は $TF\text{-}IDF^{満足} = 1{,}916$という *TF-IDF* の結果を得られた。“満足”という単語の *TF-IDF* 値が高くなった理由については、今回用いたレビューデータの評価ポイントの割合は、図 1 で示した通り、高評価であることを意味する「評価ポイント」の値が“4”または“5”のレビューの割合が 90%を占めていることから、“満足”という単語を用いたレビューが多くなったためであると考えられる。

一方で、図 3 の頻出単語上位に入っている単語(名詞)のうち“購入”、“パソコン”、“中古”、“商品”は今回のデータが EC 市場における中古ノート PC の購入者のレビューである都合上、*TF-IDF* が高くなることは自明であり、購入者の需要の分析に使うことは適していないと考えられる。従って、5.2 節以降の分析では、上記の名詞を除外した結果について考察する。

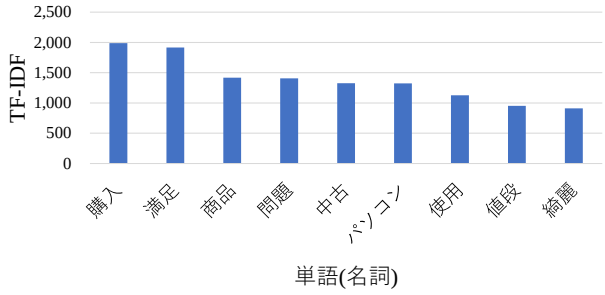
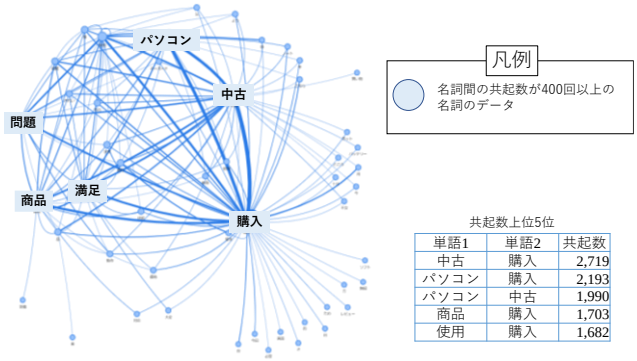


図 3. 全レビューデータにおける各単語の出現回数

Step 1.2.2 単語(名詞)の共起ネットワーク

図 4 は、「レビュー内容」の単語に対する共起ネットワーク結果を表している。図 4 より、Step 1.2.1 で *TF-IDF* が高かった“購入”、“商品”、“問題”、“中古”については、他の単語とは結びつきが高く、反対に“満足”については、*TF-IDF* が高くても他の単語とは独立していることがわかった。このことから、レビューに“満足”という単語を使う場合、つまり中古ノート PC が消費者の需要を満たしているとレビューのコメントが短くということが考えられる。



共起数上位5位

単語1	単語2	共起数
中古	購入	2,719
パソコン	購入	2,193
パソコン	中古	1,990
商品	購入	1,703
使用	購入	1,682

図 4. 全レビュー内容に対する共起ネットワーク

5.2 評価ポイントごとのテキストマイニング

5.2.1 Step 2.1 ポジティブ／ネガティブレビューデータの抽出

ポジティブレビューとネガティブレビューによって得られる各単語について重要度と共起ネットワークの違いを見るため、Step 2.1.1 と Step 2.1.2 でポジティブレビューとネガティブレビューデータの抽出を行う。

Step 2.1.1 では、5.1 節で抽出した中古ノート PC のレビューデータに対してデータ項目の「評価ポイント」が“1”かつ「レビュー内容」の評判分析結果が“NEGATIVE”のデータを抽出し、ネガティブデータとした。ネガティブデータの判別に「レビュー内容」の評判分析結果も使用したのは、購入者の中には「評価ポイント」を“1”にしていても「レビュー内容」が好意的な内容しか書かれていないなど、評価ミスと思われるデータがあった。それらは以降の分析においてノイズとなる可能性があるため、「レビュー内容」が“POSITIVE”のデータは「評価ポイント」が“1”であっても Step 2.2 の *TF-IDF* の取得、共起ネットワークでは使わないことにする。

同様の理由で、Step 2.1.2 でもデータ項目の「評価ポイント」が“5”かつ「レビュー内容」の評判分析結果が“POSITIVE”のデータを抽出し、ポジティブデータとした。

5.2.2 ポジティブ／ネガティブレビューの形態素解析

Step 2.2.1 ネガティブレビュー単語(名詞)の *TF-IDF* の算出

Step 2.2.1 では、Step 2.1.1 で抽出したネガティブレビュー(154 件)のデータ項目「レビュー内容」に対して、*TF-IDF* を算出し、その上位 10 位の単語を図 5 に示した。図 5 より、“残念”の *TF-IDF* が一番高かった。“残念”が含まれている「レビュー内容」を見てみると、中古ノート PC の性能や品質に言及しているレビューが多いことがわかった。また、*TF-IDF* が“バッテリー”に対して高いことから「レビュー内容」を見てみると、中古ノート PC で、購入したノ

ート PC のバッテリーが寿命を迎えて使えなくなっていることで購入者の不満になっていることがわかった。

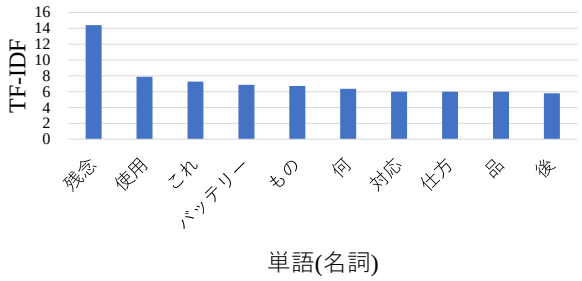


図 5. ネガティブデータにおける各単語(名詞)の *TF-IDF*

Step 2.2.2 ポジティブレビュー単語(名詞)の *TF-IDF* の算出

Step 2.2.2 では、Step 2.1.2 で抽出したポジティブレビュー(6,221 件)のデータ項目「レビュー内容」に対して、*TF-IDF* を算出し、その上位 10 位の単語を図 6 に示した。ポジティブレビューについては、“満足”という名詞の *TF-IDF* が最も高いことがわかる。また、“きれい”や“傷”の単語を含んでいる「レビュー内容」について確認していると外観が綺麗であることや、傷がないことについて言及していることから、ポジティブレビューについては、中古ノート PC の性能の次に、見た目が重視していることがわかった。

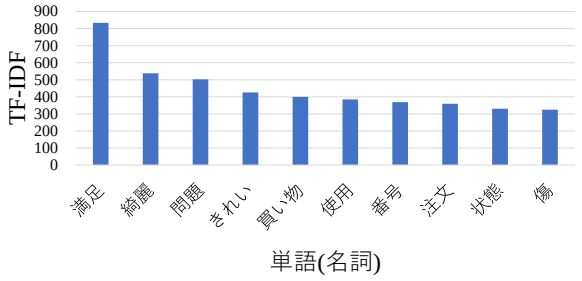


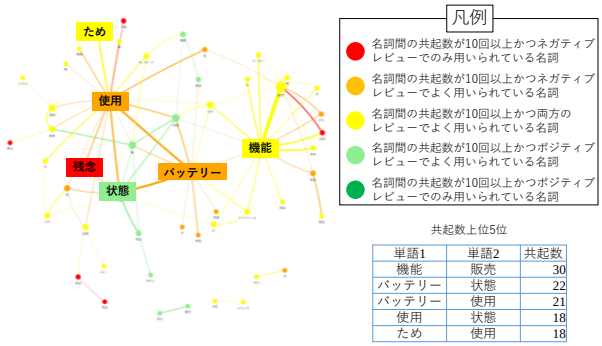
図 6. ポジティブデータにおける各単語(名詞)の *TF-IDF*

Step 2.2.3 ポジティブ／ネガティブレビューの共起ネットワーク

Step 2.2.3 では、Step 2.1 で抽出したポジティブレビューとネガティブレビューを用いた共起ネットワークについて考察する。しかし、Step 2.1 で抽出したポジティブデータ件数は、ネガティブデータ件数の 60 倍以上ある。したがって、共起ネットワークの単語のペアがポジティブレビューに偏らないために、ポジティブレビューデータはネガティブレビューと同じ件数になるように抽出した。

図 7 は、ポジティブレビューとネガティブレビューを用いた共起ネットワークである。図 7 より、“機能”という単語については、ポジティブレビューとネガティブレビューの両方(黄色)に含まれており、購入者が中古ノート PC の性能によって、評価をしていることがわかった。また、“使用”と“バッテリー”という単語に対しては、ネガティブレビューにおいて多く見られている(オレンジ)ことから、普段の使い辛さについて購入者は不満を持ちやすいことがわかった。また、“残念”という単語はネガティブレビューのみ(赤色)に見られ、購入者のレビューを満たしていないときによく使われていることがわかった。

同数のレビューで比較した場合、共起ネットワークはポジティブレビューより、ネガティブレビューで使われていた単語の共起数が多くなる傾向にあることがわかった。これは、バッテリーのように購入者の「この製品は最低限この品質を満たしているはずだ」という願望を意味する当たり前品質が満たされていても購入者のレビューでは、ほとんど言及されることがない。反対に当たり前品質を満たされていない場合に購入者はレビューで、言及するためであると考えられる。



単語1	単語2	共起数
機能	販売	30
バッテリー	状態	22
バッテリー	使用	21
使用	状態	18
ため	使用	18

図 7. 評価ポイント 1 と 5 を用いた共起ネットワーク

6. おわりに

本研究では、中古品に対するユーザの需要を明らかにするため、EC サイト上の中古品の購入者のレビューに対してテキストマイニングを行った。実験結果から、購入者は中古であっても性能が本人の需要を満たしていなければ、ネガティブレビューになること。さらに、購入者の当たり前品質を満たしているときに評価は高くなるがレビューで言及はしないこと。反対に満たされていないときにレビューは低くなり、レビューで言及する傾向がわかった。

今後の研究として、本研究のテキストマイニングでは、単語の名詞にのみ着目したが、他の品詞の単語に着目することが考えられる。さらに、ユーザの需要を満たすような単語が EC サイトの商品の説明にどれくらい入っているかを考察する。

参考文献

[1] Wang H. F., Gupta S. M., “Green Supply Chain Management: Product Life Cycle Approach”, McGraw-Hill, (2011).
[2] United Nations Environment Programme, “Global Resources Outlook 2019: Natural Resources for the Future We Want - Summary for Policymakers”, Available online: https://wedocs.unep.org/handle/20.500.11822/27518.
[3] 環境省, “リユース市場規模調査報告書”, (2022), https://www.ieice.org/publications/conference-FIT-DVDs/FIT2023/data/html/program/pdf/F-050.pdf.
[4] 我妻 幸長, “Bert 実践入門 PyTorch + Google Colaboratory で学ぶあたらしい自然言語処理技術”、翔泳社, (2023).
[5] Liu, B., “Sentiment Analysis and Subjectivity”, In Handbook of Natural Language Processing, 2nd ed.; Indurkhya, N., Damerau, F.J., Eds.; Chapman and Hall/CRC: New York, NY, USA, pp. 627–666, (2010).
[6] 赤石 雅典, “現場で使える！ Python 事前言語処理入門”, 翔泳社、(2020).
[7] Raschka, S., Mirjalili V., “[第 3 版] Python 機械学習プログラミング: 達人データサイエンティストによる理論と実践”, 株式会社インプレス, (2022).
[8] 山内 長承, “Python によるテキストマイニング”, オーム社, (2017).
[9] Zhu, D., Lu, S., Wang, M., Lin, J., Wang, Z., “Efficient Precision-Adjustable Architecture for Softmax Function in Deep Learning”, IEEE Transactions on Circuits and Systems II: Express Briefs, Vol.67, No.12, pp.3382-3386 (2020).