Study on Transition of Image Classification – from AlexNet to Vision Transformer

Juan Park, Sogang Lee, Jeongkyu Lee, Georyang Park, Jiseok Son

***Abstract*— 본 연구는 컴퓨터 비전(Computer Vision, CV)에서 이미지 분류(Image Classification) 모델의 트렌드 변화 과정을 살펴보고, 모델 성능의 효율성을 분석해 보았다. 합성곱신경망(Convolutional Neural Network, CNN) 기반 모델인 AlexNet, VGGNet, GoogLeNet, ResNet과 트랜스포머(Transformer) 모델인 Vision Transformer가 있다. AlexNet 부터 병렬GPU, MaxPooling layer, 과적합 방지를 위한Dropout, ReLU 함수, 주변 값에 영향을 줄이는LRN(Local Response Normalization)을 적극 활용해 성능 개선에 적용되었다. VGGNet은 Convolution filter가 한번에 볼 수 있는 영역의 크기인 3x3 Receptive Field가 도입되었고, 1x1 filter를 사용해 공간 정보(spatial information)를 보존할 수 있게 되었다. GoogLeNet의 경우 모델의 층(layer)을 더 깊게 쌓아도 계산효율을 증대 시킬 수 있는 Inception Module이 활용되었다. ResNet에서는 gradient Vanishing 문제를 해결 할 수 있는 Residual Network가 도입되었다. 층의 개수에 따라 ResNet18,34는 Identity(Building Block)를 사용하였고, ResNet50, 101, 152는 Projection(Bottleneck Building Block)이 적용되었다. Vision Transformer(ViT)는 이미지를 패치 단위로 잘라 토큰화 시켜준 뒤 패치에 linear embedding을 적용해 sequence 로 만들고 트랜스포머에 입력해 학습시켜준다. 대량의 데이터셋에서 학습한 경우 CNN 모델 보다 더 적은 리소스로 높은 성능을 나타냈다. 본 연구에서는 CNN 기반 딥러닝 모델의 개선 및 성능향상의 흐름을 확인 할 수 있으며, 트랜스포머 모델로의 전환의 장단점과 향후 나아가야 할 방향을 고찰 해 볼 수 있다.**

I. INTRODUCTION

딥러닝 (Deep Learning)의 발전으로 Computer Vision(CV) 분야가 급부상하게 되었고, 이미지 분류(Image Classification), 객체탐지(Object Detection), 이미지 분할(Image Segmentation), 객체 추적(Object Tracking) 등 다양한 분야에서 주목받고 있다. 딥러닝은 인공신경망(Artificial Neural Network)에 여러 개의 층을 쌓아 기계학습(Machine Learning) 하는 것을 의미한다. 대량의 데이터(이미지 또는 문자)를 기반으로 중요한 특성 및 규칙을 학습 하고 예측 가능 하게 설계되었다. Browse State Of the Art기준에 따르면, 이미지 분류(Image Classification)는 가장 많은 성능평가(Benchmark)를 갖추고 있다. 대표적으로, AlexNet은 이미지 분류에 합성곱 신경망 (Convolutional Neural Network, CNN)이 효과적 성능을 보여줄 수 있음을 보여주었다. 이후 VGGNet, GoogLeNet, ResNet으로 점차 개선되어 가면서 CNN 기법을 활용한 이미지 분류가 지배적으로 자리잡게 되었다. 하지만 Transformer 기반 이미지 분류 모델이 등장하여 CNN 분류 모델들의 성능을 능가하게 되었다. Vision Transformer의 경우 이미지 분류 Benchmark에서 최고의 성능을 나타내고 있다.

A. AlexNet

AlexNet은 CNN을 활용한 모델 중 가장 간단한 구조인 LeNet-5와 비슷한 구조를 가지고 있으며, GPU 병렬 연산을 위해 병렬 구조로 설계되었다는 점이 특징이다.

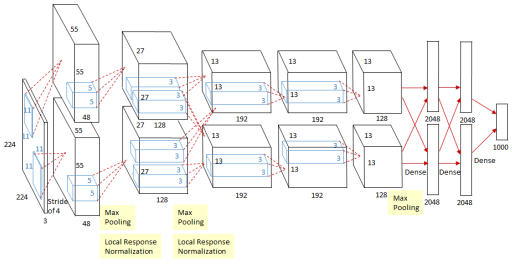


Fig. 1. AlexNet의 구조

활성화 함수로 Tanh가 아닌 ReLU를 사용하는데, 0보다 큰 값은 그대로 활성화시키기 때문에 값이 매우 큰 경우에 큰 값 주변의 지점에 영향을 끼칠 수 있다. 이 때 LRN(Local Response Normalization)을 사용하여 feature map들 사이의 동일 지점끼리 normalization을 진행하였고, 큰 값을 줄이는 방법으로 문제를 해결했다.

또한, dropout과 데이터증강(data augmentation)을 통해 과적합을 방지했다. LeNet-5에서는 average pooling을 사용하였고, stride와 kernel size를 같게 하여 pooling끼리 서로 겹치지 않게 하였는데, AlexNet에서는 stride를 kernel size보다 작게 하여 pooling끼리 서로 겹치게 하였고 max pooling을 사용하여 성능을 향상시켰다.

B. VGGNet

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Fig. 2. VGGNet의 구조

VGGNet은 layer의 깊이에 따른 성능을 확인하려는 목적을 가지고 있었기 때문에 kernel size를 3x3으로 고정하

여 층을 깊게 쌓더라도 feature map의 크기가 한 번에 축소되지 않게 하였다. 또한, 기존에 5x5, 7x7의 kernel size를 사용하는 것보다 3x3의 filter를 각각 2번, 3번 사용하는 것이 모델의 parameter 수를 더 줄여주기 때문에 깊은 층과 별개로 학습을 더 빠르게 진행할 수 있었다. (5x5의 parameter 수는 25개, 7x7의 parameter 수는 49개인 것에 비해, 3x3의 filter를 2번 사용하면 18개의 parameter, 3번 사용하면 27개의 parameter를 가지게 된다.)

Figure2와 같이, 원 논문에서는 총 6개의 VGG Model에 대해 layer 깊이를 다르게 하여 서로 비교하였다. 이 중, 16개의 layer로 구성된 VGG16(Model D)에 대해 직접 구현해보았다. 원 논문에서 Model A를 제외한 나머지 모델에 대해서는 학습 시, 학습된 Model A의 weight과 bias를 사용하여 학습하였는데, 이를 참고하여 Model A를 먼저 학습시킨 뒤, Model A와 Model D의 중복되는 layer에 대해 Model A의 weight과 bias로 Model D의 weight과 bias를 초기화하여 구현하였다.

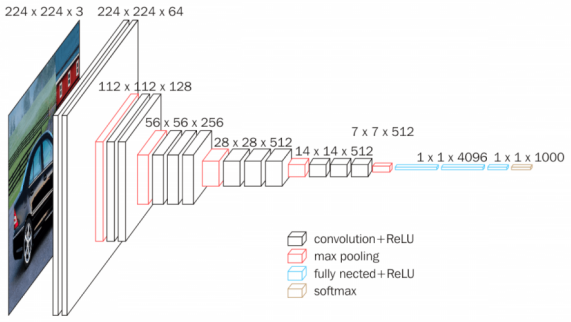


Fig. 3. VGG16의 구조

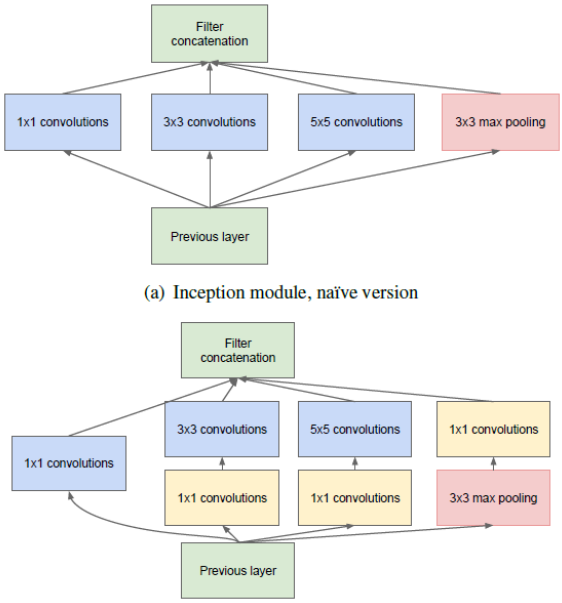
C. *GoogLeNet*

GoogLeNet의 구조는 figure 4와 같다.

Fig. 4. GoogLeNet의 구조

GoogLeNet의 특징을 크게 둘로 나눠보면, 연산량을 줄이기 위한 목적을 가진 것과 성능을 향상시키기 위한 목적을 가진 것으로 나눌 수 있다. 먼저 연산량을 줄이기 위해서 1x1 convolution과 Global Average Pooling을 사용하였다. 1x1 convolution은 feature map의 채널 수를 줄이는데 사용되는데, 1x1 convolution filter의 output channel 크기만큼 feature map의 채널 수를 줄여서 연산량을 감소시킨다. 또한, 기존에 feature map들을 flatten 해준 후, FC layer를 통과시켜서 1차원 벡터를 만든 후 softmax를 통해 분류한다면 GoogLeNet에서는 이 과정을 Global Average Pooling으로 대체하였다. Global Average Pooling은 각 feature map을 평균내어 1x1의 feature map들로 만들고 이렇게 만들어진 1x1 feature map들을 연결하여 1차원 벡터로 만드는 것이다. Global Average Pooling을 하게되면 parameter의 연산을 통해 1차원 벡터를 만드는 것이 아니라 각 feature map들의 평균만 계산하면 되기 때문에 연산량을 감소시킬 수 있다.

GoogLeNet의 성능을 향상시키기 위한 구조는 Inception module과 auxiliary classifier이다. Inception module의 형태는 figure5와 같다.



(a) Inception module, naïve version

(b) Inception module with dimensionality reduction

Fig. 5. GoogLeNet의 Inception module

Inception module은 1x1, 3x3, 5x5의 covolution과 3x3의 max pooling을 통과한 feature map들을 concatenation을 통해 합쳐서 이를 output으로 만드는 구조이다. 이 과정을 통해 서로 다른 feature가 표현된 feature map들을 하나의 output으로 사용할 수 있게 된다. GoogLeNet에서는 figure5의 (b) 구조를 사용하여 1x1 convolution을 통해 feature map의 채널 수를 줄인 후 3x3, 5x5 convolution을 통과시키거나 3x3 max pooling을 통과한 후 1x1 convolution을 통과시켜 feature map의 채널 수를 줄였다.

GoogLeNet에서 연산량을 줄이기 위해 여러 방법을 사용한 것은 GoogLeNet의 layer 깊이가 깊기 때문이다. 이 때 네트워크의 깊이가 깊어질수록 gradient vanishing, 즉 기울기 소실 문제를 해결하기 어려워진다. 역전파 계산 과정에서 초반 부분의 layer에 대해 가중치를 업데이트할 때, gradient가 점점 작아져서 0에 가까워지게 되면 가중치에 대한 훈련이 제대로 이루어지지 않아서 정상적으로

업데이트되지 않는다. GoogLeNet에서는 이러한 문제를 해결하기 위해 auxiliary classifier를 사용하여 모든 layer를 통과한 output에 대해서만 역전파를 계산하는 것이 아니라, 중간에 2개의 auxiliary classifier를 사용하여 중간 과정의 output에 대해서도 역전파를 계산할 수 있게 해준다. 중간 과정 output의 역전파 계산 과정에서는 gradient가 아직 덜 작아져서 0에 가깝지 않기 때문에 정상적으로 초반 layer에 대해서 가중치 업데이트를 진행할 수 있게 해준다.

D. ResNet

ResNet은 VGG-19를 토대로 convolution layer를 더 쌓아서 만든 모델이다.

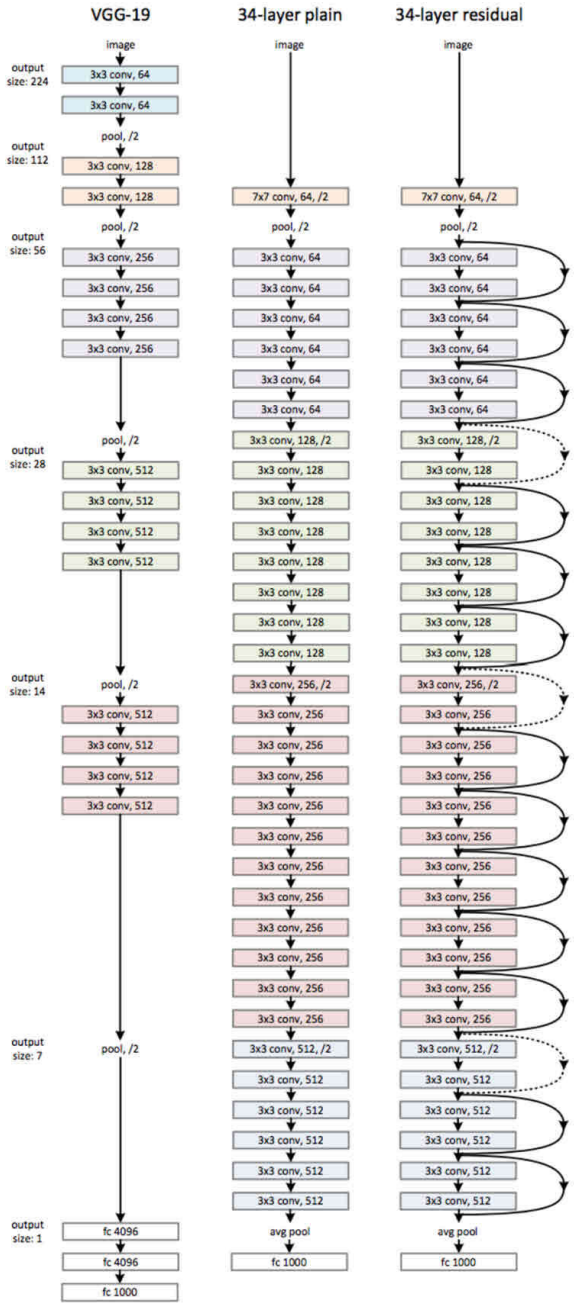


Fig. 6. ResNet의 구조

Figure6의 34-layer plain의 구조가 VGG-19에서 convolution layer를 추가한 것인데, 실제 ResNet의 구조는 34-layer residual의 형태이다. ResNet에 사용된 residual block은 layer를 깊게 쌓으면서 발생하는 gradient vanishing을 방지하는 역할을 하게 된다. Residual block은 convolution layer를 통과한 결과값에 input 값인 x를 더해주는 과정을 의미한다.

$$H(x) = F(x) + x \qquad F(x) = H(x) - x$$

위와 같은 식에서 H(x)는 input x를 넣었을 때 출력되는 결과를 의미하고, F(x)는 convolution layer를 통과한 중간 결과값을 의미한다. 이 때 H(x)를 x와 같게 해주는, 즉 F(x)를 0에 가깝게 만드는 identity mapping을 적용하고자 residual block을 사용하였고, F(x)는 곧 H(x)와 x의 차이인 잔차가 되기 때문에 residual block이라고 부르게 되었다.

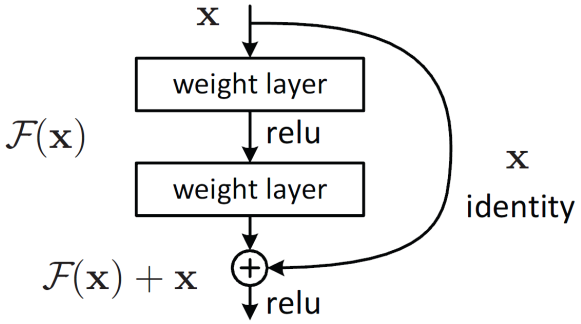


Fig. 7. Residual Block

ResNet에서 역전파를 계산할 때, H(x)에 대해 미분하여 계산하게 된다. H(x)는 F(x) + x이고 F(x) + x를 미분하면 어떤 값 + 1이 되어서 1이 항상 남게되기 때문에 gradient vanishing을 해결할 수 있다.

E. Vision Transformer

CNN 계열의 모델들이 convolution layer를 이용하여 2차원 input을 통해 이미지 분류를 수행한다면 transformer의 attention 개념 및 positional embedding을 활용하여 이미지 분류를 수행한 모델이 vision transformer이다.

II. METHOD

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. Please do not use it for A4 paper since the margin requirements for A4 papers may be different from Letter paper size.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations

III. MATH

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use

of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter", not "webers/m2". Spell out units when they appear in text: "...a few henries", not "...a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc". (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$\alpha + \beta = \chi \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is..."

D. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum ?0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

IV. USING THE TEMPLATE

Use this sample document as your LaTeX source file to create your document. Save this file as **root.tex**. You have to make sure to use the cls file that came with this distribution. If you use a different style file, you cannot expect to get required margins. Note also that when you are creating your out PDF file, the source file is only part of the equation. *Your TEX → PDF filter determines the output file size. Even if you make all the specifications to output a letter file in the source - if you filter is set to produce A4, you will only get A4 output.*

It is impossible to account for all possible situation, one would encounter using TEX. If you are using multiple TEX files you must make sure that the "MAIN" source file is called root.tex - this is particularly important if your conference is using PaperPlaza's built in TEX to PDF conversion tool.

A. Headings, etc

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and

elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

B. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
AN EXAMPLE OF A TABLE

One	Two
Three	Four

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 8. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization A[m(1)]", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K."

V. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression, "One of us (R. B. G.) thanks . . ." Instead, try "R. B. G. thanks". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
[4] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
[5] E. H. Miller, "A note on reflector arrays (Periodical styleÑAccepted for publication)," IEEE Trans. Antennas Propagat., to be publised.
[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical styleÑSubmitted for publication)," IEEE J. Quantum Electron., submitted for publication.
[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
[10] J. U. Duncombe, "Infrared navigationÑPart I: An assessment of feasibility (Periodical style)," IEEE Trans. Electron Devices, vol. ED-11, pp. 34–39, Jan. 1959.
[11] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. Neural Networks, vol. 4, pp. 570–578, July 1993.
[12] R. W. Lucky, "Automatic equalization for digital communication," Bell Syst. Tech. J., vol. 44, no. 4, pp. 547–588, Apr. 1965.
[13] S. P. Bingulac, "On the compatibility of adaptive controllers (Published Conference Proceedings style)," in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8–16.
[14] G. R. Faulhaber, "Design of service systems with priority reservation," in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 3–8.
[15] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in 1987 Proc. INTERMAG Conf., pp. 2.2-1–2.2-6.
[16] G. W. Juette and L. E. Zeffanella, "Radio noise currents n short sections on bundle conductors (Presented Conference Paper style)," presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.
[17] J. G. Kreifeldt, "An analysis of surface-detected EMG as an amplitude-modulated noise," presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.
[18] J. Williams, "Narrow-band analyzer (Thesis or Dissertation style)," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
[19] N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
[20] J. P. Wilkinson, "Nonlinear resonant circuit devices (Patent style)," U.S. Patent 3 624 12, July 16, 1990.